Project no. 2005-021324
ASPI
Audiovisual to Articulatory Speech Inversion

Special Targeted Research Project
FET

Deliverable D1
Technology inventory and specification of fields investigated

---

Due date of deliverable: November 15 2006
Actual submission date: November 15 2006

Start date of project: November 1st 2005 Duration: 36 months
Organisation name of lead contractor for this deliverable: CNRS-LTCI

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)	
Document Identifier	ASPI/2006/D1/v2.0
Revision	v2.0
Date	November 27, 2008
State	final
Dissemination level	**Public (PU)**
Authors	Shinji Maeda (LTCI) Marie-Odile Berger (LORIA) Olov Engwall (KTH) Yves Laprie (LORIA) Petros Maragos (ICCS-NTUA) Blaise Potard (LORIA) Jean Schoentgen (ULB)

ASPI Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-2005-021324.

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA)
615 rue du jardin botanique
54600 Villers-lès-Nancy
France
Contact person: Yves Laprie
E-mail address: Yves.Laprie@loria.fr

Kungl Tekniska Högskolan (KTH)
Lindstedtsvgen 24
SE-100 44 Stockholm
Sweden
Contact person: Björn Granström
E-mail address: bjorn@speech.kth.se

Laboratoire Traitement et Communication de l'Information (LTCI)
46, rue Barrault
75013 Paris
France
Contact person: Shinji Maeda
E-mail address: maeda@tsi.enst.fr

Université Libre de Bruxelles (ULB)
50, Av F.-D. Roosevelt
B-1050 Bruxelles
Belgium
Contact person: Jean Schoentgen
E-mail address: jschoent@ulb.ac.be

Institute of Communication and Computer Systems (ICCS-NTUA)
National Technical University of Athens
School of Electrical & Computer Engineering
Zografou, Athens 15773
Greece
Contact person: Petros Maragos
E-mail address: maragos@cs.ntua.gr

Contents

Chapter 1

Introduction to Audiovisual-to-articulatory inversion

Acoustic-to-articulatory inversion, or mapping, designates the recovery of either the vocal tract area function, or of the shape and position of the articulators from the speech wave. The area function is the cross-section of the vocal tract as a function of the distance from the glottis. Hereafter, the term ”shape” is used whenever one would like to refer indiscriminately to parameters that describe the vocal tract area function, or articulatory postures and gestures.

Strict acoustic-to-articulatory inversion requires that the exact vocal tract shape that produced the observed acoustic data is recovered, rather than one of the many possible tract morphologies that can produce the same acoustic data [1]. Many researchers in acoustic-to-articulatory inversion have instead focused on estimating plausible tract shapes, that is, cross-sections or postures that are compatible with human speech production. The distinction between actual and admissible tract shapes stems from the fact that the computation of the tract cross-sections from acoustic data is a problem that may admit more than one solution.

The causes of this non-unicity are multiple. They include

- the existence of qualitatively different area functions that are characterized by the same eigenfrequencies, when the wave propagation is loss-less [1];
- that the number of unknown morphological output parameters may exceed the available acoustic input parameters;
- the inability to exactly recover duct shapes that reproduce the acoustic data.

The non-unicity problem differs in severity, however, according to whether the inversion involves (i) the transfer function of the vocal tract, or only the formant frequencies;(ii) a model of the human vocal tract, or an unconstrained area function and (iii) a loss-less or a lossy propagation of the acoustic wave. Generally speaking, the non-unicity problem is worst when the area function must be recovered from measured formant frequencies, and the wave propagation is assumed to be lossless. For this case, Mermelstein [2] has shown that the area function is not unique, whatever the number of the formant frequencies that are given.

A unique solution of the acoustic-to-area conversion problem, however, exists under the following conditions [3]:

- The glottal excitation is a single unit pulse;
- the area function is a concatenation of identical cylindrical tubelets with arbitrarily large cross-sections;
- the total length of the vocal tract is known a priori and equal to an integer multiple of the lengths of the individual cylinders;
- all losses are resistive and concentrated at the glottal or labial end of the tract.

The concatenated-cylinder area function can then be determined by means of a linear regression model that is fitted to the noise-free vocal tract impulse response. The condition is that the order of the model is equal to the number of cylinders, which must be known a priori. Obviously, human vocal tracts do not fulfill the assumptions of loss-less wave propagation or single-unit pulse excitation. Also, the tract length is not known a priori. As a consequence, area functions thus estimated from recorded speech signals are considered to be rough approximations at best, the anatomical plausibility of which is not guaranteed [4].

In practice, experimenters must decide on (i) the source of the data to build the knowledge about the vocal tract shapes for different phonemes; (ii) on the model of the vocal tract that is used for the inversion; (iii) the representation of the acoustic data; (iv) the inversion method; (v) the constraints to use in order to alleviate the non-uniqueness of the mapping and finally (vi) how to evaluate the performance. The aim of the following chapters of this inventory is to describe how these choices may be made.

The **articulatory data** of the vocal tract can be collected using numerous techniques, and since no technique is currently able to capture all aspects required for speech inversion, several techniques *must* be used, as described in Chapter 2.

The **vocal tract models** that have been used range from articulatory sagittal-profile models to unconstrained area function models. Sagittal profile models either include mimics of human articulators or the principal components obtained by a statistical analysis of the sagittal profiles of a human speaker. A heuristic is then used to turn the two-dimensional profile into a three-dimensional duct, through which plane acoustic waves propagate. The different types of vocal tract representations that can be and have been applied to articulatory inversion are described in Chapter 3.

The **acoustic data** that is the input to the inversion method may be represented as formant frequencies or whole-spectrum features. Examples of the latter are cepstral and linear-predictive-coding coefficients. Formant frequencies offer a phonetically meaningful description of vowel-like speech sounds. Formant-to-shape conversion has therefore been the preferred option of those who advocate computational inverse mapping with a view to the study of the link between speech signals and tract shapes. The different representations of acoustic data are described in Chapter 4.

Acoustic-to-articulatory inversion algorithms that have been proposed may be divided into several categories. The main distinction is between model-based and statistical methods. Model-based approaches, which use little articulatory data, were ruling until roughly 25

years ago. When the computerized handling of large amounts of data became feasible and computation-intensive data compression methods were developed that enabled the data to be represented economically (e.g. [5] [6]), the the drift has been continual from model-based data-poor to model-free data-rich methods in all areas of speech processing, including acoustic-to-shape mapping.

In principle, the availability of large amounts of articulatory data recorded via imaging of the vocal tract of a speaking subject, in conjunction with the corresponding acoustic data, would enable discarding models altogether. Indeed, the combined acoustic and articulatory data can be organized either in the framework of a codebook or compressed further by means of deterministic or stochastic learning methods.

In practice, however, the amount of natural training data has often been so small that natural data had to be augmented or replaced by synthetic training data that are generated by means of a vocal tract model. That is, articulatory or area function models are used to produce morphological-acoustic data pairs that are organized in the framework of codebooks or compressed by means of automatic learning algorithms, as if the data were natural [7]. In one study, the synthetic data was replaced by virtual data [6]. That is, from a set of acoustic data, equivalent tract-shapes were recovered computationally and the resulting acoustic-geometric data pairs compressed by means of artificial neural nets, as if the data were genuine.

Data-free methods rely on models exclusively. They fall into two categories according to whether they use models in a forward (i.e. causal) or inverse direction. When used forwardly, the model parameters are manipulated iteratively until the synthetic acoustic output agree with the observed acoustic data. That is, the acoustic-to-shape transform is turned into a problem of optimization or control. The alternative consists in inverting explicitly the link between morphological and acoustic data in the framework of a model and use the ”backward” model as a new model, the input of which is acoustic and the output morphologic.

One question relevant to the project at hand is whether the importance taken by data-driven methods has lead to a genuine increase in performance. Given the lack of comparative studies, the cursory evaluation of existing inverse transforms, and the scarceness of genuine data, this question has no final answer yet.

Often, data-rich and data-free methods have been used in cascade. That is, raw or compressed codebooks have been used to find tract shapes that roughly reproduce the observed acoustic data; in a second stage these are refined by means of a data-free method until the calculated and observed acoustic data agree as far as possible. Data-free methods have also been used to discover the set of admissible tract morphologies that coexist with a single tract-shape solution found by codebook lookup [5].

The different types of algorithms that may be employed for acoustic-to-articulatory inversion are described in Chapter 5.

Constraints that need to be applied in acoustic-to-articulatory inversion are of two different types. The role of the first type of constraints is to assure that the output represents articulatory configurations that are anatomically possible for a speaker. The second type of constraints is needed to attempt to find which articulatory position that actually produced the input speech sound, since it can been shown that anatomically possible articulatory postures do not guarantee a unique solution [8]. Articulatory trade-offs are expected to exist in the case of speech

sounds that involve double articulations (e.g. rounded vowels). That is, quasi-identical speech sounds can be produced with tract shapes that differ qualitatively. Examples of a human use of that capacity are ventriloquist and bite-bloc or lip-tube speech.

Additional constraints are therefore necessary. In practice, these have been wide and varied. Restrictions on tract shapes, which have been used in isolation or in combination include the following:

- Overdetermination, that is, the number of acoustic cues exceeds the number of morphological parameters [9];
- Expansion of the log-area function by means of a small number of odd Fourier cosine coefficients [2] [10];
- Imposing maximal and minimal tract cross-sections [11];
- Minimization of the distance between the recovered and a neutral tract shape [12];
- Maximization of the spatial smoothness of the computed area function [12];
- Keeping the vocal tract volume constant [13];
- Maximization of the smoothness of the temporal evolution of the shape parameters [14];
- Minimization of the temporal rate of change of the shape parameters [14];
- Minimization of ”muscle” work [15];
- Matching of output and radiation impedance [10];
- Maximization of radiated acoustic power [10].

These constraints permits to select a single solution among all the possible solutions that are anatomically acceptable for a given speech sound. Generally speaking, however, the constraints are intuitively satisfying at best, but most lack a rigorous justification on the base of human speech production.

It is indeed so that speaking is a secondary function of an apparatus the primary function of which is breathing or biting, chewing and swallowing. The forces that are involved in the latter are larger than the forces involved in speech production. One may therefore wonder whether the minimization of distance, speed or acceleration are valid criteria for speech articulators that are controlled by muscle forces that are feeble compared to those that are applied when chewing or biting, for instance. Also, the movement of human limbs in general is not subject to the restriction that the acceleration is minimal. The constraint rather is that the rate of change of the acceleration (i.e. jerk) is small. A constraint of minimal jerk appears to apply to articulatory movement [16].

Another viable type of constraint is to use available information of the speaker’s visible articulators (lips and jaw) to limit the possible configurations of the invisible ones (tongue, velum, larynx) [17–19]. This introduction of visual information, which transforms the problem into

audiovisual-to-articulatory inversion is a key focus of the ASPI project. Constraints in general, and facial information constraints in particular, are described in Chapter 5.

Evaluation of the performance of the mapping is a central concern to most studies in speech-to-shape mapping, since they have mostly dealt with the inversion per se. The authors disagree on the evaluation criteria, however. The reason is that different applications request different criteria of performance. The main distinction appears to be between usages that explicitly request the computed shapes to be veridical because users expect to learn about human speech production as such. Examples are in speech therapy, second language learning, aids to the handicapped as well as computational imaging for phonetic purposes. Such evaluations have been scarce and in practice, the evaluation of the truthfulness of computed shapes has been quite elementary. Some studies have involved static vowel shapes, mainly Fant's Russian vowels, to judge if the recovered vocal tract shapes are plausible [2, 9, 10, 12, 15, 20–25]. A minority of studies have involved metallic, rubber or numerical models.

A majority of applications focus on acoustic-to-tract mapping not as a substitute imaging technique, but as a non-trivial transform of acoustic into morphological data. The expected benefit is that the morphological data possess qualities that acoustic data lack, such as smooth and slow evolution, or the property of further decomposability into constituents that are context-independent. These are desirable properties in the framework of articulatory synthesis, speech compression as well as automatic speech or speaker recognition. The criteria of success here are that the speech-shape-speech transform is consistent and the evolution of the morphological data smooth. Veridical reproductions of the tract shapes are of minor or no importance. As a consequence, evaluation has not focused on the genuineness of the recovered shapes, but on the agreement between observed and modeled acoustic data, or on informal listening to resynthesized speech sounds. When the objective was automatic speech recognition, the evaluation has involved rates of correct recognition.

The **possible applications** of successful acoustic-to-articulatory are numerous, but the state-of-the-art in acoustic-to-articulatory is still one of research and development. That is, only a small minority of published articles have reported on acoustic-to-tract mapping as an instrument, the performance of which is taken for granted. Examples of studies that make use of acoustic-to-area mapping as a tool are [26] [27] [28]. [27] reports on a device that is commercially available. The duct shape, however, is not reclaimed from the speech signal, but from the impulse response of the nasal and pharyngeal tracts.

A large majority of publications focus on acoustic-to-shape inversion as a topic of research or experiment. To our knowledge, no tool that has been based on speech-to-shape inversion has gone beyond demonstrator status. None has been widely accepted as industrial, clinical or laboratory implement. Putative applications that have been cited, but not necessarily implemented, are:

- Visual representation of speech for the totally or partially deaf [29];
- Speech therapy; investigation of articulatory control in speech disorders [15] [11];
- Articulatory feedback for second language learning [11] [7];
- Investigation of physiological processes involved in speech production, and of the link between speech signal and vocal tract shape [5] [30];

- Replacement of imaging by computation in phonetic investigations to avoid exposure to X-rays and the laborious task of (manual) image processing [7] [21] [31];
- Investigation of the challenge to the concept of ”place of articulation” posed by the ability to produce the same low-frequency formants by different tract shapes [8];
- Investigation of the recovery by humans or machines of linguistic units from speech [31] [32];
- Speech compression [33] [11];
- Improving speech quality of lossy coders by means of an articulatory model by imposing smoothness constraints on the area function and its motion [20] [7] [14];
- Articulatory synthesis and low-bit rate synthesis of speech that would include tract-source interaction [5] [34] [35];
- Facilitation of the control and training of speech synthesizers because of the ability to interpolate articulatory trajectories, and the easy timing of transitions between speech segments as well as of the onset and offset of turbulence noise [36] [37] [11];
- Facilitation of the design of rules for articulatory synthesis [9];
- Linear decomposition of morphological data further into components that are context-independent [29] [38];
- Feature extraction for automatic speech recognition [5] [11] [39];
- Exploitation of the location of critical articulators to discard some hypotheses in the framework of automatic speech recognition [7];
- Speech segmentation based on articulatory postures and movement [40];
- Establishment of a link between automatic speech recognition and speech synthesis [41];
- Automatic adaptation of a speech recognizer to the speaker via the adjustment of the total tract length [38].

Chapter 2

Imaging & measurement of speakers' vocal tract and face

The development of new speech inversion methods and their evaluation is tightly connected to the availability of tools used for the imaging of the vocal tract. Thanks to technological advances, a wide variety of such tools are now available to measure the shape and movements of the vocal tract, with increasing detail and accuracy. Historically, many techniques have been used to gain knowledge about the vocal tract, including plaster casts of living or dead subjects, fibrescobe filming, Computed Tomography, electropalatography, optopalatography etc. This survey of measurement methods will concentrate on X-ray, X-ray microbeam, Ultrasound, electromagnetic tracking and Magnetic Resonance Imaging, which are the most relevant for the data required within the ASPI project. In addition, techniques for collecting data of the speaker's face, using optical tracking or video-based methods will be described, since the use of visual constraints derived from the speaker's face may provide important information to solve the acoustic-to-articulatory inversion problem.

Ideally, the technique used to measure the movements of the vocal tract should:
(i) Cover the whole vocal tract with all the articulators and the face visible.
(ii) Give a time resolution sufficient for the tracking of the dynamics of the vocal tract.
(iii) Not involve any known health hazard for subjects,
(iv) Not perturb the natural articulation
(v) Not degrade the quality of the speech signal recorded together with images and
(vi) Be portable for field works outside laboratory.

Currently, no such method exists; all methods require the subject to produce the articulations under more or less unnatural circumstances and no method measures the full 3D geometry and kinematics at the same time, as discussed below and summarized in Tab. 2.1. The purpose of this survey is to inventory the techniques that could be used to collect acoustic-articulatory data that are necessary for the development and evaluation of new speech inversion methods.

2.1 Vocal tract

2.1.1 Cineradiography, X-ray

Cineradiography has traditionally been the main information source on real-time movement in the midsagittal plane, since it was used for the first time in the 1920's ([42]). The advantage of X-ray imaging is that it provides real-time measurements of the entire two-dimensional tongue contour in upright position. Modern digital X-ray equipment makes it possible to record 50 images per second with a spatial resolution of about 0.3 mm. A difficulty with X-rays is to accurately identify the vocal tract structures in the images, as discussed in Section 2.4 and exemplified in Fig. 2.1. To enhance the contrast in the images, subjects swallow a viscous liquid that adheres to the tongue surface, to the floor of the mouth and to the lips. The accuracy of the contours traced is of the order of 0.5 - 1 mm.

The importance of X-ray measurements in speech research is indicated by the extensive bibliography compiled by Dart [43], listing 282 X-ray studies, done in a large number of languages, including, e.g., the influential studies by Fant [44–46].

The use of X-ray measurements has however been drastically reduced and restricted over the last decades, as the hazards for the subjects became apparent. Recent developments in digital X-ray technology have permitted to minimize the subjects' exposure to radiation, by using a prespecified pediatric program, by removing the image intensifier scattered-radiation grid and by placing the subject's eyes outside the imaging area. The absorbed dose in the most exposed organ (parotis sin) is less than 4 mGy during a 20 second acquisition, with the effective dose not exceeding 0.1 mSv (i.e. an amount corresponding to a tenth of the annual natural background radiation), as in [47]. This means that X-ray measurements are viable again for small corpora that can be used for assessment of a model or to measure specific details of

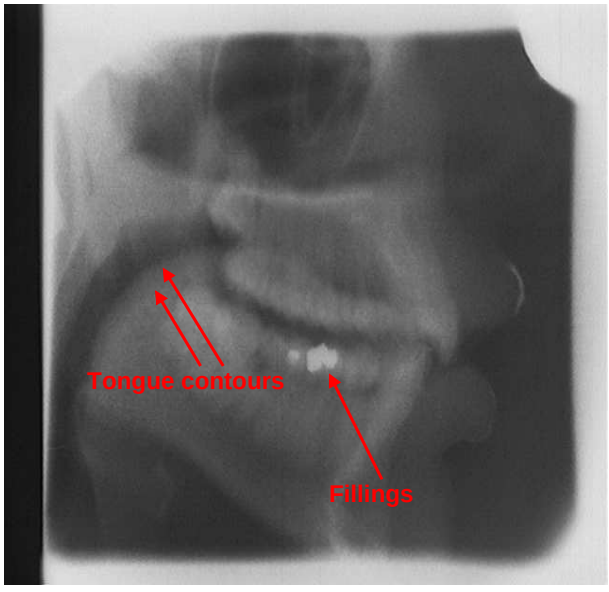


Figure 2.1: *Example of X-ray image from the Munhall et al. database (film 55)*

an articulation. Other measurements are however required as a basis for modeling, where a larger corpus is required.

2.1.2 X-ray microbeam

The ethical constraints of exposing subjects to X-rays for non-medical purposes still remain, and the radiation dose has to be limited as much as possible. A possibility is to use the X-ray microbeam technique [48, 49]. In X-ray microbeam, a very narrow beam of high-energy x-rays is generated, and rapidly directed under high-speed computer control, to track the motions of 2-3 mm diameter gold or lead pellets glued to the tongue, jaw, lips, and soft palate. The main advantage over traditional X-ray measurement, apart from the reduction in the radiation dose for the subject, is that the amount of data is reduced from a continuous shadow to clearly defined discrete points, facilitating the data processing. This of course means that information on the remaining contour is lost, but it can to some extent be reconstructed through interpolation and combination with other data sources, such as articulatory models based on X-rays [50] or principal component analysis [51]. The X-ray microbeam technique has hence been successful in characterizing anterior tongue movements, but its inherent limitations prevent imaging of the tongue root and pharynx, which have important effects on the acoustic output. Other important limitations of this technique, including the fact that it is rather expensive, invasive and not portable, have restricted its use in speech research.

2.1.3 Electromagnetic articulography (EMA)

Another point-wise midsagittal measurement method is Electromagnetic Articulography (EMA), which employs alternating magnetic fields instead of X-rays. EMA tracks midsagittal fleshpoints movements by measuring induced current from receiver sensors moving in a magnetic field [52] [53]. The magnetic field, generated at different frequencies by two to six transmitter coils, induces an alternating signal in the receiver coils. Since the voltage of this signal is inversely related to the distance between the transmitter and the receiver coil, a computer algorithm can determine the location of the receiver coils as they move in space, based on the voltage.

Six receiver coils are commonly used for the measurements: two are placed on the upper and lower lip, three coils on the tongue (approximately 8 mm, 20 mm and 52 mm from the tip of the tongue, depending on speaker) and one on the base of low front incisors (to measure jaw movement). A typical placement of EMA coils is shown in Fig. 2.6(a). In addition, two receiver sensors, one on the base of the upper front incisor and one on bridge of the nose, are often used as reference points for head-movement correction. Rotation and translation of the EMA sensor data is performed to ensure that the two reference coils are coincident across all frames for a given speaker. This removes any component of head movement from the data. A further rotation is performed to align the occlusal plane (also called “bite plane”) with the x-axis and a translation sets the origin at the position of the upper incisor reference coil.

The most important advantage of EMA is the high tracking rate, with sampling of articulatory data at 200 Hz. Another advantage is related to the ability of tracking multiple articulators simultaneously. These two aspects make it possible to measure, with increasing accuracy, co-

ordination among different articulators. Notice, however, that the accuracy of the data recorded decreases away from the center of the triangle of the transmitter coils. Hoole [53], for instance, reported an error of 0.67 mm +/- 0.42 for positions more than 6 cm away from the center (in the midsagittal plane) and 0.2 mm +/- 0.13 for positions up to 6 cm. Another important aspect concerning the reliability of EMA data is related to rotational misalignments. Articulatory data can only be collected on the midsagittal plane and are thus subject to error as the articulators rotate left-to-right [53].

Different EMA systems are available, including the the MIT system [52], the Botronic Movetrack system [54], and Carstens Articulograph (http://www.articulograph.de). Carstens AG100 is by far the most used system among speech researchers. It is comprised of (1) a plastic helmet that subjects wear during data recordings (three transmitter coils are mounted equidistant from one another on the helmet) (2) small receiver coils placed inside the mouth or on the face and (3) an electronic unit connected to the computer.

One interference with the subject's natural speech is the transmitter coil helmet and the more recent AG500 [55] overcomes this by replacing the helmet by a 'cage' on which six transmitter coils are fastened. AG500 hence allows for free head movements and the sensors can further be positioned outside the midsagittal plane and in all orientations. This means that the EMA measurements have gone from being two-dimensional and point-wise to three-dimensional, but still point-wise. The three-dimensional EMA system is still very much under development and midsagittal EMA remains the standard articulography method. Two-dimensional EMA has been used in a large number of studies to explore tongue movements [56–58].

2.1.4 Ultrasound echograph

Ultrasound can be used for either kinematic two-dimensional (at 30-200 Hz) or static three-dimensional measurements. The technique employs a transducer probe containing piezoelectric crystals, that change shape rapidly when subjected to an electric current. As the crystals vibrate, high-frequency (5-40 MHz) sound waves are emitted, and conversely, when a sound wave is absorbed by the crystal, it emits an electric current. This current can be used to reconstruct a wedge-shaped image of the midsagittal slice of the tongue, as shown in Fig. 2.2.

The sound waves are reflected against boundaries where there is an important change in density. This means that only the outer tongue body shape can be measured as the available boundary is that between the tissue and the air. Parts where there is also air underneath, such as the tongue tip, when it is lifted, or the palate, do not show up. The measurements are hence often restricted to the tongue body as the tongue root is obscured by the hyoid bone.

A good introduction to the ultrasound technique, its theoretical principles and properties, is given in [59], which was one of the first suggestions for using ultrasound in speech research. Methodological and technical questions in using ultrasound for tongue measurements are also addressed in [60], regarding the system setup, transducer placement and aspects such as peak detection and measurement resolution. Ultrasound has been since been used in a number of speech production studies, focused on two-dimensional cross-sectional movements e.g., [61–64] or the three-dimensional tongue shape [65, 65]. In the first type of study, the probed is held in a fixed orientation during running speech, and in the latter, it is moved to different

orientations during sustained phonation. The technique is currently in particular advocated by Stone and colleagues [66, 67], who have designed both measurement set-ups and data analysis software.

Data collection using Ultrasound is well suited for the imaging of the vocal tract and offers many advantages. It involves no health risks for the subject. It offers a relatively high temporal resolution: slices can be collected at video rates (30fps) for analogue systems and over 100 fps for digital ultrasound echographs. It is relatively inexpensive and portable, and subjects are recorded in a natural, upright speaking situation. The lack of tongue tip data in the ultrasound images can also be compensated for by using an electro-magnetic tracking system. The shape of the hard palate can also be obtained by asking the subject to take a mouthful of water and force it up into contact with the hard palate. Since the impedance difference between water and tongue tissue is different from that between water and bone, the palate shape can be extracted. However, to subsequently inserted the palate shape into the ultrasound images of the speech production measurements requires that the head position is known during the experiment.

This issue of defining an absolute spatial reference in the signal is central in ultrasound imaging. Knowing where the ultrasound image is in relation to the vocal tract is a very difficult problem that must be solved. If one is interested in the shape of the tongue only, then measurements in a jaw-centered system (i.e. the probe is allowed to move with the jaw) is sufficient. To determine the exact location of the tongue within the vocal tract, however requires that the position of the jaw is also known, One way to do this is to immobilize the head and the probe using a specially designed system, such as the Head And Transducer Support HATS [67]. The subject's head is fixated in the HATS system and the transducer position is adjusted until the best image possible of the tongue is obtained.

The recently developed Haskins Optically Corrected Ultrasound System (HOCUS) [68] does not require immobilization. The system incorporates both ultrasound imaging of the tongue and optical tracking of the position of the ultrasound probe relative to the head. The optical system (Optotrak) tracks the location of external structures on the head and on the ultrasound probe in three-dimensional space using infrared emitting diodes. The head, probe, and jaw are allowed to move, but their motion is tracked and can therefore be used to correct the tongue measure-

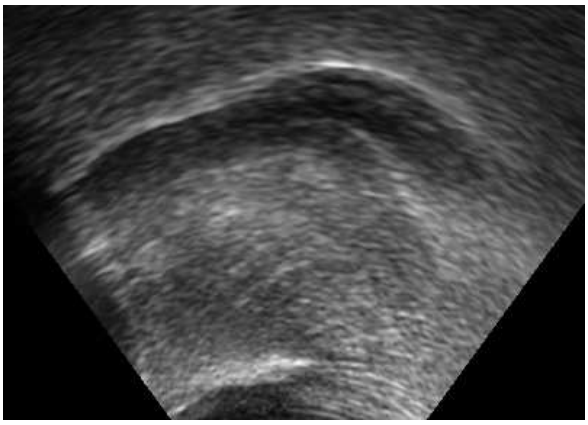


Figure 2.2: *Ultrasound image of the tongue, with the tip to the left.*

ment to a head-based coordinate frame. Different visible structures (such as the lips and jaw) can also be tracked and complete measurements of the vocal tract during fairly unconstrained speech are hence recorded. Stone and colleagues have developed software for the analysis of ultrasound images, such as EdgeTrak, an automatic system for the extraction and tracking of tongue contours [69] [70]. A few points on the tongue in the first image are chosen, and EdgeTrak then uses an active contour model to determine the location of the tongue edge in the current and following images. Though this contour tracking system, currently used by scientists in several institutions, is accurate enough for speech research, it still has some weaknesses. For instance, since the ultrasound images are quite noisy and there are some unrelated high contrast edges in the images, the gradient information is sometimes insufficient to extract edges of interest. The tongue contour may also be interrupted in places.

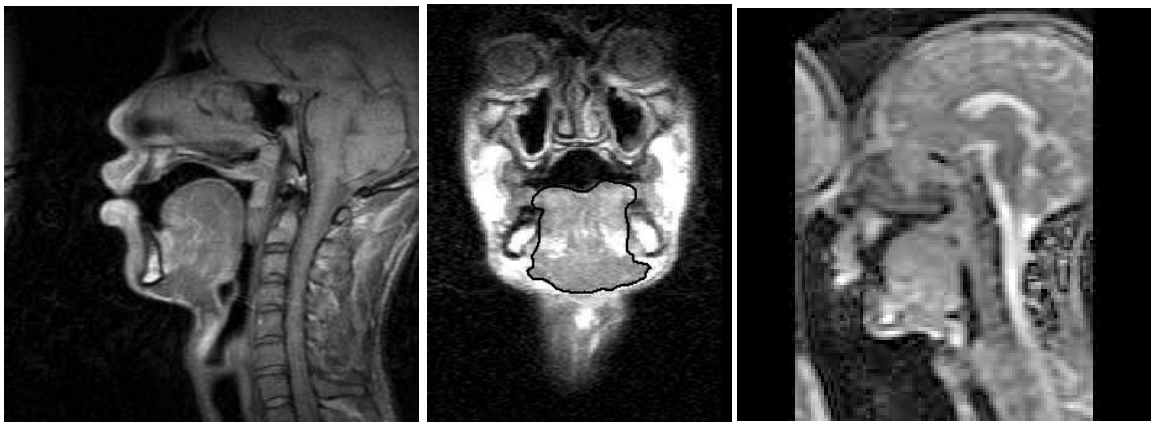
2.1.5 Static Magnetic Resonance Imaging

Since Magnetic resonance imaging (MRI) was first used to analyze the vocal tract [71, 72], it has grown to be the dominating method for measuring speech production three-dimensionally in many different languages. The success during the past decades is based on image features and quality and relative subject-friendliness

The basis for MRI is that the hydrogen atoms in the body can be aligned using a strong induced magnetic field. A radio frequency pulse is directed towards the area of the body that is to be examined and the proton of the hydrogen atoms absorbs energy that makes it spin in a different direction. Using pulses of a specific frequency, the Larmour frequency, the protons can be made to precess in a determined direction. Once the pulse is turned off, the protons return to their natural alignment in the magnetic field, and in doing so they release the surplus energy, which can be captured by the magnetic coil. The data of the energy release can then be converted into a picture using Fourier transforms. Using gradient magnets that are turned on and off very rapidly, the magnetic field can be altered in a small area, which means that MRI is able to collect data in slices of 2-5 mm at any orientation, as exemplified in Fig. 2.3(a)-(b). These features allow two-dimensional images of (approximately) two-dimensional arbitrarily oriented slices to be collected and combined into detailed 3D images.

The images have good signal to noise ratio, are amenable to computerized 3D modeling, and provide excellent structural differentiation. In addition, the tract (airway) area and volume can be directly calculated. MRI is also subject-friendly in the sense that it has no known harmful side effects and no ethical constraints need hence to be put on the amount of data that can be collected. The technique has however several disadvantages for speech production measurements, such as that the electromagnets produce high amplitude noise. The noise is caused by the rising electrical current in the wires of the gradient magnets being opposed by the main magnetic field and its amplitude is proportional to the strength of the main field. The noise make simultaneous acoustic recordings difficult, although not impossible, if optical microphones are used.

The most severe disadvantage of the technique is the prolonged acquisition time, during which the subject must remain immobile, as even slight movements of the scanned body part cause distorted images. When MRI first was used for speech measurements [73], 30 minutes were required to obtain the full set of images for a given vocal tract configuration and the



(a) Midsagittal (b) Coronal (c) Midsagittal real-time

Figure 2.3: *MR Images at different orientations (a-b) and with different acquisition times.*

subjects had to produce a sustained monotone for the 3.4 minutes it took to acquire each image, breathing in briefly every 15 seconds. The technical advances, that allow the acquisition of the entire vocal tract to be made with high image quality in around 30 seconds, is hence a very important contributing factor to the success of the method. The acquisition times needed are still decreasing and it is now possible to collect full 3D data in 5 seconds. However, the images collected with such short acquisition times are often of low quality, and the standard use of MRI in speech production measurements is for the study of artificially sustained speech sounds. This sustaining may cause the articulations to be both hyperarticulated and more difficult to hold for the subject [74]. The results may be a backward (i.e. a downward in the supine position) movement of the tongue and a lack of velum control, as exemplified by the blurring caused by movement of the velum in Fig. 2.3(a). It is hence important to keep the MRI acquisition time as short as possible.

Another disadvantage is that the acquisition is made with the subject in supine position, which may affect the articulation. Several MRI studies, including [75] [74], have noted a backward displacement of the tongue caused by the supine position. Tiede et al. [76] found postural effects between sitting and lying position in X-ray microbeam measurements and Engwall [74] showed that the vocal tract became more constricted in the pharynx when the subject was facing upwards, as opposed to downwards, in the MRI acquisition. The static MRI images thus have to be complemented with other measurements (e.g. EMMA, EPG, or X-ray) to correctly replicate not only articulatory movements but also positions in running speech.

2.1.6 Dynamic MRI

The acquisition times for MRI have decreased drastically during the past years and methods to image the moving vocal tract are emerging. One possibility is to use many repetitions of a phoneme string and generate a real-time image sequence through post-processing [77–81]. As the articulation varies slightly between the repetitions, the reconstructed image sequence show an aggregate of all the repeated articulations, rather than the true articulation, and this

may lead to discontinuities in vocal tract shape over the sequence.

The large number of repetitions may introduce variability in the articulations and the development of sensitive encoding systems or ultra fast Turbo Spin Echo, allowing to capture several (4-24) images per second [82–84], is hence a great advance in dynamic MRI, as real-time capturing of slowly produced sequences can be made with the technique. The MRI technique is hence approaching a time resolution where many articulatory movements can be studied in real time. The limit at 100 Hz, which is often considered as the lower limit for visual synthesis, is however still far away, and other real-time measurements techniques will probably still be required. The image quality in real-time MR imaging is further far inferior to imaging of sustained articulations with longer acquisition times, as shown by the comparison of Figs. 2.3 (a) and (c).

2.1.7 Summary of measurement techniques

For the sake of clarity, the different techniques described above are compared in Table 2.1.

	EMA	**MRI**	**Ultrasound**	**X-ray**	**X-ray microbeam**
Time resolution[1]	200 Hz	0-24 Hz	30-200 Hz	50 Hz	40-160 Hz
Whole V.T.	No	Yes	No	yes	No
Tongue imaging	Pellets	Full-length	Full-length	Full-length	Pellets
Tongue root	No	Yes	No	Yes	No
Velum imaging	Yes[2]	Yes	No	Yes	Yes
3D	No	Yes	No	No	No
Health hazard	No	No	No	Yes	Yes
Natural art.	Affected	Yes[3]	Yes	Yes	Affected
Acoustic noise	Low	High	Acceptable	Low	Acceptable
Head Mvt.	Restricted[4]	Restricted	Restricted[5]	Free	Free
Portable	No	No	Yes	No	No
Inexpensive	No	No	Yes	No	No

Table 2.1: Comparison of vocal-tract shape recording techniques. Notes: [1] It is sufficient to have 60 frames/s to observe muscular-force induced articulatory movements, while 1000 frames/s would be required to observe aerodynamic-force induced movements, such as those during consonantal release. [2]To record velum position data using EMA or X-ray microbeam, a receiver or a pellet have to be attached to the velum. [3]The supine position during MRI recording may affect the articulation. [4]Head movement is free using 3D EMA and restricted using a 2D system. [5]As was mentioned in text, head movement is free using the Haskins Optically Corrected Ultrasound System (HOCUS).

Due to the limitations of the above measurements techniques, no single one is able to fulfill the requirements for acoustic-articulatory measurements. A combination of MRI for three-dimensional static measurements and combined ultrasound and electromagnetic tracking for tongue movements will therefore be employed. In addition, already existing X-ray, X-ray microbeam and EMA data, available in research databases will be explored to provide additional insights.

2.2 Face

Measurements of the speaker's face are important for articulatory inversion for two reasons. Firstly, they give information about visible articulators that are directly involved in speech production, i.e. the jaw and lips. Secondly, they can provide indirect information about articulators that are invisible from the outside view through statistical relationships between the position of these articulators and the appearance of the face.

The data of the face are of two types, shapes and movements. Shape data is most commonly collected using laser scanning or structured light range digitizers, to achieve a dense map of the face. Movement data may be collected using either special optical motion tracking systems or video images (from one or two cameras). In order to evaluate the anatomical changes occurring during speech, knowledge of the shape and size of the speaking face is required. While several systems have been already implemented that allow $2D$ visual features to be automatically extracted from a video sequence [19, 85] showing a talking face the extraction of $3D$ visual features that could improve speech inversion still remains unexplored.

Modeling and measurements of the human face have wide applications ranging from medical purposes [86, 87] to computer animation [88–90], from video surveillance to lip reading systems, from video teleconferencing to virtual reality [91–94]. The issues that must be considered to model the face of a real person are: how realistic and accurate the obtained shape is, how long it takes to get the result, how simple the equipment is and how much it costs.

2.2.1 Face shape measurements

To date, the most popular measurement technique is ***laser scanning*** [87, 95], for example the head scanner of Cyberware [96]. This systems normally scans the human face in about 30 seconds. The subject has to remain still while the scan platform moves a digitizing unit around the head. The digitizer is composed of a light beam and video cameras to capture all details of the object, colours included. With triangulation or interferometry methods, 3-D coordinates of the scanned points can be quickly computed. These systems give a dense cloud of measured points and are easy to use. The achieved accuracy is limited to 0.5 mm and smooth filters have to be applied on the modeled surface because of its roughness. The long scanning period contributes to the low precision as the subject cannot remain absolute immobile for so long. These scanners are expensive and the data are usually noisy, requiring touch-ups by hand and sometimes manual registration.

Another solution is offered by the ***structured light range digitizers*** [97, 98] which are usually composed of a stripe projector and one or more CCD cameras. Several products based on this technology are available (see www.eietronics.com, www.inspeck.com).

Such a system is usually composed of a camera and a programmable projector. Defined sequences of stripe images are projected onto the object during the acquisition. A time-space coding of a sequence of n stripe images allows the differentiation of $2n$ different projection directions. Given a calibrated projector and camera, the depth information can be computed through triangulation using the acquired images. The system is simple to use and is practical to install (only one camera and a projector). For these reasons it has gained importance in the

industrial sector. The method is optimal for static objects. For complex objects such as the human face, multiple acquisitions from different directions and with different projection directions are required. These can be used for face reconstruction with relatively inexpensive equipment compared to laser scanners.

The ***accuracy*** of both systems is satisfactory for static objects. However, their acquisition time ranges from a couple of seconds to half of a minute, depending on the size of the surface to measure. Thus, a person must remain stationary during the measurement. Not only does this place a burden on the subject, but it is also difficult to obtain stable measurement results. In fact, even when the acquisition time is short, the person moves slightly unconsciously.

A different approach to face modeling uses images as source data. Various ***image-based techniques*** have been developed. They can be distinguished by the type of used image data: a single photograph, two orthogonal photographs, a set of images, video sequences or multi-images acquired simultaneously.

Parametric face modeling techniques [88, 99] start from a single photograph to generate a complete 3-D model of the face. Exploiting the statistics of a large data set of 3-D face scans, the face model is built by applying pattern classification methods. The results are impressively realistic, however the accuracy of the reconstructed shape is low.

A number of researchers have proposed creating models from two ***orthogonal views*** [100]. Manual intervention is required for the modeling process by selecting feature points in the images. It is basically a simplified method to produce realistic models of human faces. The obtained shape does however not reproduce the real face precisely. To solve this problem, some solutions [89] work in combination with range data acquired by laser scanners.

The ***multiple view based methods*** is another image-based method, which consists of automatically extracting the contour of the head from a set of images acquired around the person [101, 102]. The obtained data are combined to form a volumetric model of the head. The set of images can be generated moving a single camera around the head or having the camera fixed and the face turning. The systems are fast and completely automatic, however the accuracy of the method is low.

2.2.2 Optical tracking systems: Qualisys and Optotrak

The Qualisys system (`http://www.qualisys.se`) consists of 1 to 16 cameras, each emitting a beam of infrared light. Small reflective markers are placed on the speaker's face; a typical placement is shown in Fig. 2.6(b). the camera flashes infra-red light and the markers reflect it back to the camera. The system is able to register the 3D-coordinates for each marker at a frame-rate of 60 to 1000 Hz.

Thirty or more markers can be used to register lip movements as well as other facial movements such as chin, cheek, eyebrows and eyelids. Additional markers may be placed on the chest to register head movements with respect to the torso. Subjects also wear a pair of spectacles with four markers attached, used as a reference to be able to factor out head and body movements when looking at the facial movements specifically.

Optotrak, which is an improved version of the older Watsmart, uses three line charged

coupled device sensors that track target points defined by up to 512 miniature Infrared Emitting Diodes. Each one of the three sensors consists of a cylindrical lens system and a signal processing circuitry. All three measurements together determine the 3D location of the infrared LED marker, which is calculated and displayed in real time. In speech research data, these target markers are used to track the motion of the jaw, lips, cheeks, and eyebrows. Additional markers, attached to a head rig, are used to define the head-based coordinate system.

The accuracy of Optotrak in the operating volume at 2 m is 0.1 mm in the x- and y- dimensions, and 0.15 mm in depth, while the resolution is 0.01 mm. The maximum sampling rate (marker frequency) is of 3500 Hz. This system is thus well suited to tracking lip and jaw motion in 3D and examining relationship between them and head position.

The disadvantage of these two point tracking systems, is that only fleshpoints are tracked, not the entire surface or volume. For rigid structures, such as the jaw, the entire structure can however easily be reconstructed knowing the exact shape.

2.2.3 Video-based tracking

2.2.3.1 Monovision

Monovision video recordings, i.e., using a single camera, can be explored in mainly two angles, either in profile or a full front view, with the latter being by far the most common for automatic analysis of articulatory features. Most existing methods for extracting information from face video rely on tracking the lip contours, which can be modeled using snake-like methods [103] [104] or data driven principal component analysis (PCA) methods [105] [106] [107].

An alternative is to track the face as a whole and then extract the mouth region of the image. The advantage is that this method is more robust, as the face is less deformable and therefore easier to track. Once the lip region has been extracted, it is possible to either find the lip contours [108] or represent the information of the mouth in terms of e.g., independent components of the lip image [109] or binary articulatory mouth features [110]. Binary features are robust for separation between a small set of words, but the independent components are more suitable for inversion, as they include important information, such as shading indicating lip protrusion, and visibility of the teeth and tongue, that is not present in the lip shape. Kjellstrom et al. [109] indeed showed that the addition of monovideo images of the speaker's face improved the automatic estimation of the positions of four electromagnetic articulography coils placed on the jaw and tongue with about 25% compared to the audio-only case. The inversion was performed on 63 Swedish VCV words using a relevance vector machine (RVM), a non-linear kernel-based regression technique.

2.2.3.2 Stereovision

Stereovision techniques allow to capture shape deformations in rapidly moving scenes. Given two (or more) calibrated cameras and given two corresponding points in the images from the two different cameras (i.e points that correspond to the same physical point), the 3D points are built as the intersection of the two rays that pass though the optical centers and the im-

age points. The key problem to be solved in stereovision is the matching stage where points that correspond to the same physical 3D point must be identified in the two images based on similar intensity or color. This problem is highly difficult and is often solved by imposing strong assumptions on the scene such as lambertian material.

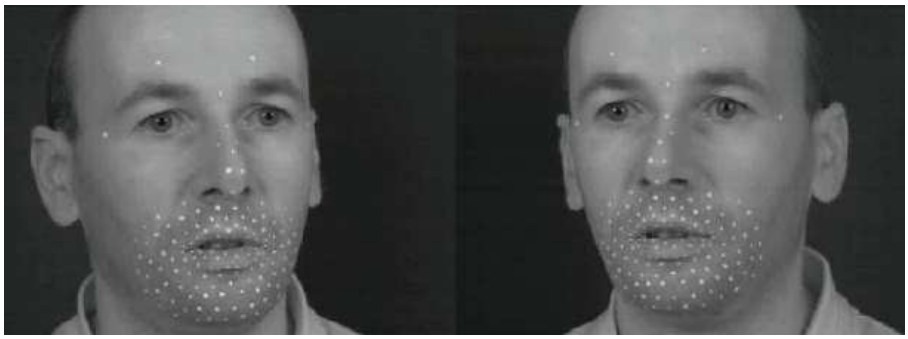


Figure 2.4: *Paired left/right images for stereovision.*

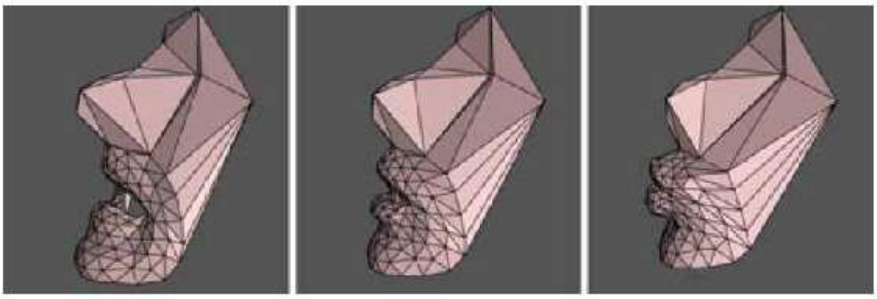


Figure 2.5: *The first three principal components of the 3D reconstructed face.*

In order to make the matching stage easier, markers can be glued or painted on the face as shown in Figure 2.4. Similarly to optical tracking, this method only allows a sparse map of the face to be obtained and is thus of limited interest for face reconstruction/modeling/synthesis. However, the method allows to get temporal 3D reconstruction of the face and is widely used to build dynamic models of the face with principal component analysis [111] [112] [113] [114]. As an example, Figure 2.5 illustrates the first three principal components of reconstructed 3D face images.

Obtaining a dense map of the face from stereovision techniques is much more difficult. Correlation techniques are widely used and have been successful in many studies. However, these techniques use a fixed neighborhood to compare the intensity in the two images whereas a surface patch may have different shapes in the two images due to projective effects. This limitation may be alleviated by considering adaptive windows [115]. In addition, correlation can only cope with affine changes of image intensities. As a result, classical reconstruction methods are not robust and false matchings are often present, especially at depth discontinuities and in regions presenting near uniform texture. Considering the matching stage as a global optimization problem has recently lead to significant improvements in stereo reconstruction.

The stereo problem is then considered as the minimization of a cost functional that integrates statistical similarity criteria and regularization constraints on the disparity map [116]. Despite promising results, it still remains difficult to obtain automatically a dense reconstruction of a speaking face, especially because imposing regularization constraints on the face movement is difficult.

For these reasons, projected light patterns are often used, that display patterns on the face to facilitate the recovery of a dense map (c.f. section 2.2.1).

A key limitation to the methods presented in section 2.2.1 is that they methods do not capture motion, i.e. a point correspondences over time, making difficult to learn the dynamic of the face. To overcome this problem, two main solutions have been developed:

- A morphable linear model of the face can be derived from a set of 3D face models. Then, reconstruction for a particular talker can be achieved by fitting 2D featured (texture) or 3D features (issued from stereovision) against a generic parametric morphable model [90, 93, 94, 117–119]. Reconstruction is then considered as a recognition problem. The techniques are fully automatic but may perform poorly on face with unusual features or other significant deviations from the normal.

- Using time varying structured light patterns projected on a face, algorithms that integrate space and time consistency in a global optimization approach significantly improve the quality and stability of the recovered depth map. In [120] [121], space time consistency is achieved by assuming that disparity is nearly constant over a 3D space time window.

Despite continuous progress, obtaining dense maps of human faces with temporal coherence using cameras is still a challenging problem. Many methods still need manual and often tedious interaction processes. One of the fundamental challenges is to reduce (or to suppress) manual intervention. The other challenge is to to have a cheap acquisition set-up and simplified acquisition techniques, especially to reduce the calibration burden.

2.3 Existing Databases

Another important resource for speech production research is the existing databases of previously collected articulatory measurements. The information contained in the following databases will be of great benefit to improve the knowledge of the acoustic-to-articulatory mapping and to improve the existing acoustic/articulatory inversion methods.

2.3.1 X-ray movie films

Due to ethical concerns about the high radiation dosages necessary, X-ray imaging technology is currently rarely employed. Therefore, it has become imperative to preserve those films that were originally captured on the fragile medium of 35 mm film. This is addressed by the X-Ray Film Database for Speech Research Project and by the IPS X-ray Database of Strasbourg.

The X-Ray Film Database for Speech Research Project collaborative project by Dr. K.G. Munhall (Queen's University) and Drs. E. Vatikiotis-Bateson and Y. Tohkura (ATR Human Information Processing Laboratories, Kyoto, Japan) was conceived to create a database that stores a collection of high-quality copies of the original x-ray films in a durable format on a constant angular velocity (CAV) format videodisc [122]. The aim is to make these images available to the speech research community free of charge and to develop techniques for automated digital processing of these images. The videodisc is available to researchers by contacting Dr. Kevin Munhall (munhallk@psyc.queensu.ca).

This database contains a series of 25 X-ray movies for a total of 55 minutes, together with the sound recordings. 24 of the 25 films are from the Université Laval and show 9 native speakers of Canadian French and 5 native speakers of Canadian English reading phonetically contrastive sentences [123]. The films have a temporal resolution of 50 frames per second, and do not show the lower pharynx or larynx, but the hyoid bone is visible and the lips and velum are clear in most of the 24 films. The 25th film was recorded at MIT and shows the entire vocal tract and lips at 45 frames per second [124].

The Institut de Phonétique de Strasbourg has gathered more than 50 X-ray recordings, including data from a large variety of languages, since the 1950's. Researchers from the Institut de Phonétique de Strasbourg, with the collaboration of the Institut de la Communication Parlée de Grenoble, have recently undertaken the task of creating a database that stores this collection of the original X-ray films in high-quality copies [125]. The aim is store these films in a durable format and make them available for speech research community.

The database currently contains 4 movies that present over 2000 images. The X-ray data focus on different phonetic issues in French: juncture, nasals, and coarticulation in VCV sequences. The database contains 3 kinds of digitized data: the cineradiographic data, acoustic signals and hand-drawn sagittal contours of the vocal tract.

All files are phonetically labeled and stored on CD ROMs. The film database is available to researchers by contacting the Institut de Phonétique de Strasbourg, the owner of the Database.

2.3.2 X-ray microbeam, University of Wisconsin

The database from the University of Wisconsin X-ray microbeam facility covers a relatively large number of speakers (about 200 different speakers), and a rich, uniform inventory of utterances and oral motor tasks, yielding a data-set more than 3200 tracking minutes. Speakers contributing to the database project were young adults from the campus of the University of Wisconsin-Madison and surrounding cities. A majority of these speakers spoke an Upper Midwest dialect of American English.

The database resource is intended to be sufficiently accurate and deep to withstand statistical scrutiny of variance, within and across speakers. This firstly requires that the task list is sufficiently broad to encompass most of the range of motor and linguistic tasks a speaker performs when talking. Secondly, the list must be sufficiently redundant to provide meaningful estimates of intra-speaker variability, and thereby allow reliable inferences regarding speaker intent, and control principles governing the speech act. Each speaker dataset contains: reading of two prose passages (13%); counting and digit sequences (6%); oral motor tasks (8%);

citation words, near-words, sounds and sound sequences (33%) and sentences (40%). The sentences consist of 21 TIMIT sentences and 19 other sentences with varying numbers of repetitions.

Eleven pellets were used, attached to the head (3 pellets, as reference markers), upper and lower lips (1 pellet each), tongue surface (4 pellets), and mandible (2 pellets). Pellets were glued to the tongue using Ketac, and to all other surfaces using Isodent (commercially-available dental adhesives), and then anchored by light threads taped to the skin surface of the cheeks and face.

Three types of time-series data were recorded for each speaker: (1) wide-band physiological tracks; (2) videophotographic images, and, (3) low-band pellet position tracks. The wide-band physiological tracks recorded the radiated sound pressure wave and a representation of neck wall vibration overlying the thyroid lamina. The video images were recorded at 60 frames per second to monitor the speakers' positions in the microbeam image field, and can be useful for understanding certain speaker movements that affect data accuracy. The image quality is not high, however, and it is unlikely that the images are useful to extract visual information of the speaker's face. The spatial resolution of the pellet tracks is inversely proportional to the distance of the image plane from the system pinhole. For database speakers, that distance was typically on the order of 53 cm.

Each record subdirectory contains a collection of files representing the time series data. In the typical case, 19 files are produced: one each for the sound pressure wave and neck wall vibration; one each for the x- and y-coordinates of each of eight articulator pellets (i.e., 16 files); and one representing a vector of explicit time stamps for the sixteen pellet-coordinate histories [126]. The database is an open resource, available for unlimited inspection by other speech scientists.

2.3.3 Multi-CHannel Articulatory database, University of Edinburgh

The MOCHA (Multi-CHannel Articulatory) database collected at Queen Margaret University College, was developed to provide a resource for training speaker-independent continuous Automatic Speech Recognition systems and for general co-articulatory studies. The publicly available database (at `http://www.cstr.ed.ac.uk/artic/mocha.html`) consists of simultaneous acoustic-articulatory data of 2 speakers of British English who read 460 phonetically rich sentences, providing data on more than 10,000 phones per speaker [127]. The 460 sentences comprise the 450 TIMIT sentences, designed to provide good phone pair coverage, along with an extra 10 sentences which include phonetic pairs and contexts found in the Received Pronunciation (RP) accent of British English. These were designed to include the main connected speech processes in English (e.g., assimilations and weak forms) used for both training and testing.

The MOCHA database offers a number of different parallel streams of acoustic and articulatory data. The acoustic data was recorded directly onto disk in a sound-damped studio sampled at 16kHz and stored with 16-bit precision. The articulatory information includes electromagnetic articulograph (EMA), laryngograph, and electropalatograph (EPG) data. The EMA data consists of x- and y-coordinates in the mid-sagittal plane for 7 points on the articulators,

sampled every 2 ms. The EMA sensors were attached to the upper and lower lips, lower incisor (jaw), tongue tip (5-10 mm from the tip), tongue blade (approximately 2-3 cm posterior to the tongue tip sensor), tongue dorsum (approximately 2-3 cm posterior to the tongue blade sensor) and velum. Subjects were screened for their ability to tolerate soft palate touching prior to the recordings. The Laryngograph measures changes in the contact area of the vocal folds, providing pitch and voiced/voiceless information and was sampled at 16kHz. The electropalatograph (EPG) provides information on tongue-palate contact at 62 normalized positions on the hard palate, sampled every 5 ms. This includes lateral tongue contact information that is missing from the EMA data.

In addition a SVHS video of front view of mouth area is also available.

2.3.4 Qualisys-Movetrack database, KTH

The Qualisys-Movetrack database at KTH consists of simultaneous recordings of the audio signal, electromagnetic articulography (EMA) data of the tongue and optical tracked facial data [128]. The recordings were made in a sound-proofed room, using a DAT tape recorder, the electromagnetic articulograph Movetrack [54] for the tongue movements and the stereo-motion capture system MacReflex from Qualisys for the face.

Six EMA receiver coils were used in the acquisition, as shown in Fig. 2.6(a). Three of them were placed on the tongue, approximately 8, 20 and 52 mm from the tip, and one on the jaw. The EMA coils on the upper lip and upper incisor were used to align the Qualisys and Movetrack data sets. Fig. 2.6(b) shows the placement of the 25 small reflectors in the Qualisys system that were used to capture the subject's facial movements. An additional 3 markers were glued to the Movetrack headmount to be able to adjust for head movements.

The subject was a female speaker of Swedish, judged as highly intelligible by hearing-impaired listeners. The corpus consisted of 270 Swedish sentences, 138 VCV and VCCCV words and 41 asymmetric C_1VC_2 words. The Swedish everyday sentences have been developed for audio-visual speech perception tests. They are independent of each other and

(a) Movetrack (b) Qualisys

Figure 2.6: *Placement of EMA coils and Qualisys markers in the KTH Qualisys-Movetrack database.*

generally seven to nine syllables long (4-5 words). The VCV words were consonants and consonant clusters in symmetric cardinal vowel context, while the CVC words consisted of long and short vowels in words with C=[p,k,r].

The speech signal was sampled at 16 kHz, the EMA data at 200 Hz and the Qualisys data at 60 Hz. To temporally align all data sets, the acoustics and the EMA data was resampled to 60 Hz. The EMA data was aligned with the corresponding Qualisys lip and jaw markers to create a coherent 3D data set of face and tongue movements.

The database contains both EMA data and parameter trajectories for a three-dimensional tongue model, estimated from the EMA data. The 3D tongue model was derived from a statistical analysis of three-dimensional MR images of one Swedish subject producing a corpus of 13 vowels in isolation and 10 consonants in three symmetric VCV contexts [58]. The estimation of the parameter trajectories for the jaw height (JH), dorsum raise (TD), body raise (TB), tip raise (TT), tip advance (TA) and tongue width (TW) is described in detail in [128].

To summarize, the estimation is based on a fitting procedure to minimize an error function that takes the goodness of fit between the model and data, the difference between the reference tongue volume and that generated by the parameter values and the parameter ranges into account. The goodness of fit was calculated as the absolute difference between the positions of the real EMA coils and the corresponding virtual coils in the model. The benefit of employing tongue model parameters rather than EMA coil positions is that the parameters have an articulatory relevance and qualitative investigations may hence be made on the relationship between articulations and acoustics.

2.4 Vocal tract image processing techniques

The processing and analysis of image data of the speaker's vocal tract is very important for the overall success of the project. For example, X-ray databases are an unequaled source of information on articulatory movements, thanks to the good time resolution (between 25 and 50 fps) and the coverage of the entire vocal tract shape. The X-ray movies in the databases described above represent a large amount of data that cannot be processed by hand. It is thus interesting to develop techniques that could partially or totally process these data automatically. Due to the difficulty of this task, few techniques have yet been proposed, despite the potential interest. The difficulties lie in the image quality of analogue X-ray films and in the nature of X-ray images that make them difficult to be processed even by a human expert, as shown in Fig. 2.1. A first problem is that the tongue contour (the most important articulator) is often hidden by teeth or dental fillings that are opaque to X-rays. A second is that since X-rays cross the head from one side to the other, several contours will be superimposed in the two-dimensional image. The tongue is thus not always represented by one contour in the midsagital plane, but by several, as the tongue edges also appear. It should also be noted that the film that the image of Fig. 2.1 comes from is of high quality compared to other films in the database.

The desired accurate recovery of vocal tract shapes can be effectively done using *image segmentation* techniques. The efficiency of the segmentation can be significantly improved by applying some appropriate pre-processing steps, such as image smoothing and interpolation: *Image smoothing* is needed to remove the noise and to simplify the image, eliminating the

objects at certain scales (sizes). *Image interpolation* may also be needed, if the subsequent analysis steps require a better image resolution than the one available (e.g. in the case of 3D MR images or for the combination of data from MR and US images having different resolutions).

As we will present in the following, all the above tasks (image smoothing, interpolation and segmentation) can be effectively approached using modern nonlinear multiscale techniques from Image Analysis and Computer Vision based on Partial Differential Equations (PDEs) and Mathematical Morphology. The use of PDEs introduces continuous models and offers better and more intuitive mathematical modeling, connections with physics, better approximation to the Euclidean geometry of the problem as well as high accuracy and stability of the corresponding numerical algorithms. In addition, the nonlinear image smoothing and segmentation can be also done using techniques of Mathematical Morphology, on which we comment briefly at the end of this section.

2.4.1 Image Smoothing

Multiscale image analysis [129] has proved to be very useful in many computer vision applications. The first scale-spaces were linear and generated using Gaussian convolutions. As Koenderink observed [130], such a Gaussian scale-space can be modeled via the homogeneous heat diffusion PDE (with initial condition the input image and the artificial time playing the role of scale parameter). This approach has been significantly improved by various nonlinear modifications of the heat diffusion PDE, so that the diffusion respects the semantically important image features. We will present some of the most important nonlinear diffusion PDE methods. Perona and Malik [131] proposed the following nonlinear heat diffusion PDE:

$$\frac{\partial u}{\partial t} = \operatorname{div}\left(g(\|\nabla u\|)\nabla u\right) \tag{2.1}$$

where $\|\nabla u\|$ is a simple edge-strength measure and the diffusivity function $g(r)$ is smooth and decreasing, with $g(0) = 1$ and $g(r \to +\infty) = 0$. With such a choice of $g(r)$, the diffusion favors intraregion over interregion smoothing (i.e. the diffusion is reduced in strong edges). For instance, a typical choice of g is $g(r) = 1/(1 + (r/K)^2)$, where K is an appropriate constant. Two problems with the diffusion scheme (2.1) are the amplification of noise by the gradient and sensitivity to initial conditions for certain choices of g. An improved model which overcomes these problems was given by Alvarez et al. [132] and its general form is:

$$\frac{\partial u}{\partial t} = g(\|\nabla G_\sigma * u\|)\left((1 - h(\|\nabla u\|))\,\Delta u + h(\|\nabla u\|)\,\|\nabla u\|\,\operatorname{div}\frac{\nabla u}{\|\nabla u\|}\right) \tag{2.2}$$

where G_σ is an isotropic 2D Gaussian of standard deviation σ, $g(r)$ is a function with same properties as before and $h(r)$ is a smooth nondecreasing function, with $h(r) = 0$, if $r \leq e$ and $h(r) = 1$, if $r \geq 2e$, where e is an appropriate constant. Thus, away from image edges (where $\|\nabla u\|$ is small), the diffusion is strong and isotropic, whereas near the edges, the diffusion is reduced and smooths the level lines of the image. PDE methods can also arise from a variational framework, by evolving the input image so that it minimizes a properly designed functional. The most popular and well-studied functional is the *Total Variation* (TV), proposed by Rudin et al. [133]:

$$TV[u] = \iint_\Omega \|\nabla u(x, y)\|\,\mathrm{d}x\mathrm{d}y \tag{2.3}$$

The minimization of this functional leads to a PDE of the form (2.1), with $g(r) = 1/r$. Minimizing the TV functional does not penalize discontinuities, but only strong oscillations, therefore the noise can be removed without blurring the edges. On the other hand, this model oversmooths homogeneous regions and destroys parts of edges in images with significant noise. In some more sophisticated PDE methods (e.g. [134]), the diffusion is not only nonlinear but also anisotropic, i.e. it is driven by an image dependent anisotropic tensor. Tschumperlé and Deriche [135] recently proposed one of the most effective methods of this kind. The method is designed for the general case of vector-valued images and can be described by the following set of coupled PDEs:

$$\frac{\partial u_m}{\partial t} = \mathrm{trace}\Big(T\big(J_\rho(\nabla \boldsymbol{u})\big) \cdot D^2 u_m\Big),\ m = 1, .., N \tag{2.4}$$

where N is the number of vector components of the image and $D^2 u_m$ is the Hessian matrix of the vector component u_m. Also, T is the *diffusion tensor*, given by:

$$T\big(J_\rho(\nabla \boldsymbol{u})\big) = \frac{1}{\sqrt{1+\mathcal{N}^2}}\, \boldsymbol{w}_- \boldsymbol{w}_-^{\mathrm{T}} + \frac{1}{1+\mathcal{N}^2}\, \boldsymbol{w}_+ \boldsymbol{w}_+^{\mathrm{T}} \tag{2.5}$$

where $\mathcal{N} = \sqrt{\lambda_+ + \lambda_-}$, with $\lambda_- \leq \lambda_+$ and $\boldsymbol{w}_-, \boldsymbol{w}_+$ being respectively the eigenvalues and the unit eigenvectors of the *structure tensor*:

$$J_\rho(\nabla \boldsymbol{u}) = G_\rho * \sum_{m=1}^{N} \nabla u_m\ (\nabla u_m)^{\mathrm{T}} \tag{2.6}$$

The convolution with G_ρ is done so that $J_\rho(\nabla \boldsymbol{u})$ takes also into account the neighborhood of every point. The eigenvectors $\boldsymbol{w}_-$ and $\boldsymbol{w}_+$ describe the orientation of minimum and maximum vectorial variation of $\boldsymbol{u}$ and the eigenvalues λ_- and λ_+ describe measures of these variations (the term $\mathcal{N}$ is an edge-strength predictor which effectively generalizes the norm $\|\nabla u\|$). Thus, the diffusion is strong and isotropic in homogeneous regions (small $\mathcal{N}$), but weak and mainly oriented by image structures near the edges (big $\mathcal{N}$). Consequently, this method offers a more reliable measure of local image variations as well as a more flexible and effective control on the diffusion process. In terms of the medical images of the vocal tract which concern us here, this method can efficiently apply the desired smoothing. It can remove the noise (even if the input image is very noisy) but also maintain and enhance the boundaries of the vocal tract.

2.4.2 Image Interpolation

PDEs have been recently used also for the interpolation of images, leading to methods which overcome the limitations of classical interpolators. Following a variational framework, Guichard and Malgouyres [136] formulated the interpolation as an inverse problem. The continuous solution of interpolation $u(x, y)$ is constrained to satisfy the following *reversibility condition*:

$$z[i, j] = Q(s * u) \tag{2.7}$$

where $z[i, j]$ is the discrete input image, $Q(\cdot)$ is a sampling operator and $s(x, y)$ is an a priori chosen smoothing kernel (e.g. the "mean kernel"). The condition (2.7) means that the decimation of the interpolation solution $u(x, y)$ should lead to the input image $z[i, j]$. The problem of

finding $u(x, y)$ in (2.7) is ill-posed and the set of its solutions forms a linear subspace (which will be noted as $\mathcal{U}_{z,s}$). Thus, the authors propose to choose as solution the image that minimizes the TV (2.3) inside this subspace $\mathcal{U}_{z,s}$. This minimization can be solved using a constrained gradient descent, which leads to the following PDE:

$$\frac{\partial u}{\partial t} = P_{\mathcal{U}_{0,s}} \left\{ \mathrm{div} \left(\frac{\nabla u}{\|\nabla u\|} \right) \right\} \tag{2.8}$$

supplemented with the initial condition that $u(x, y, 0)$ is given by the zero-padding interpolation (i.e. the element of $\mathcal{U}_{z,s}$ whose Fourier transform takes zero values outside the baseband). $P_{\mathcal{U}_{0,s}}\{\cdot\}$ is the operator of orthogonal projection on the subspace $\mathcal{U}_{0,s}$. This method leads to reconstructed images without blurring effects (as it allows discontinuities) and preserves one-dimensional image structures, such as edges. On the other hand, it tends to oversmooth homogeneous areas and it cannot always avoid some artifacts, such as staircase effects. Belahmidi and Guichard [137] have further improved the above approach by developing a non-linear anisotropic PDE, which performs adaptive smoothing quite similar to the Tschumperlé's model (2.4). They also took into account the reversibility condition (2.7), adding to the PDE a reaction term so that the flow $u(x, y, t)$ stays "close" to the subspace $\mathcal{U}_{z,s}$. This method balances between linear zooming on homogeneous regions and anisotropic diffusion near edges, combining the advantages of these two processes. Therefore, it leads to reconstructed edges without blurring and with natural shape but also does not oversmooth homogeneous areas.

2.4.3 Image Segmentation

2.4.3.1 Snakes

Snakes or active contour models were proposed by Kass et al. [138] and are one of the most important tools in computer vision. Active contours are based on deforming a given contour enclosing an initial region, until it comes into alignment with the boundary of the image object to be segmented. This is accomplished by minimizing a properly designed energy functional of the contour, which should be at (local) minimum when the contour delineates the desired object boundaries. Regularization of the contour is achieved by additional energy terms which favor the smoothness of the curve and limit the bending effect. The energy minimization is accomplished by steepest descent techniques, which lead to curve evolution governed by a PDE. This model has been effectively used in many applications and leads to a quite fast algorithm. On the other hand, the classical approach of snakes cannot directly deal with changes in topology, i.e. even when needed, the evolving contour(s) cannot split or merge. Special ad hoc procedures have been developed to overcome this limitation, but they are complicated and heuristic. In addition, the snake functional is not intrinsic, as it depends on the arbitrary curve parametrization. This is an undesirable property, since parameterizations are not related to the curve geometry.

One of the first attempts to use Snakes to extract tongue contours was that of Tiede [139]. At first sight, the method is well suited for this purpose, but because other features are superimposed on the tongue, tracking cannot be achieved alone with snake methods. Berger and Laprie [140, 141] thus proposed a tracking tool that combines Snakes and a motion based

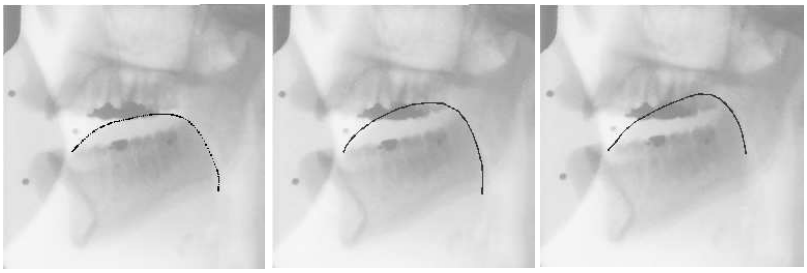


Figure 2.7: *Example of tongue tracking in three consecutive frames using a combination of Snakes and a motion based method.*

method, in which the motion is estimated through the calculation of optical flow. Fig. 2.7 shows an example of tracking for three consecutive images.

The operator outlines the tongue in the first image and supplies a rectangle containing a motionless region (from the top of the image up to the dental fillings in the upper jaw). This region is used during a pre-processing stage necessary to remove spurious image movements and intensity variations between images.

2.4.3.2 Geodesic Active Contours

Caselles et. al [142] proposed *Geometric Active Contour* models to overcome the drawbacks of classical Snakes. In the method, the PDE of the geometric curve evolution was directly designed from a heuristic point of view. These models further developed in a variational framework by Caselles et al. [143] and Kichenassamy et al. [144], leading to the *Geodesic Active Contours* (GAC). In this model, the curve evolution is derived from the minimization of an energy functional, which is a geodesic in a Riemannian manifold, endowed with a metric induced by image features. Therefore, in contrast to snake energy, this functional is not depending on the curve parameters. This geodesic (weighted length) to be minimized is given by:

$$J[\vec{C}] = \int_0^{\text{Len}(\vec{C})} g(||\nabla I(\vec{C}(s))||)ds = \int_0^1 \underbrace{g(||\nabla I(\vec{C}(q))||)}_{\text{edge attraction}} \underbrace{||\vec{C}'(q)||}_{\text{smoothness}} dq \tag{2.9}$$

where $\vec{C}(q) = (x(q), y(q)), q \in [0, 1]$ is a curve, s is the Euclidean arc length parameter, $\text{Len}(\vec{C})$ is the Euclidean length of $\vec{C}(q)$ and g is a smooth decreasing function s.t. $g(0) = 1$, $g(r \to \infty) = 0$. For instance, a typical choice of g is $g(r) = 1/(1 + r^n)$, with $n = 1$ or 2. The minimization of this energy using steepest descent leads to a curve evolution. After the addition of a "balloon" force to speed up convergence, this evolution finally becomes:

$$\frac{\partial \vec{C}(t)}{\partial t} = \underbrace{(g\kappa)\vec{N}_i}_{\text{smoothness}} - \underbrace{(\nabla g \cdot \vec{N}_i)\vec{N}_i}_{\text{edge attraction}} - \underbrace{(g\beta)\vec{N}_i}_{\text{balloon force}} \tag{2.10}$$

where κ is the curvature of the contour, $\vec{N}_i$ is the unit inward normal to the contour and β is a constant specifying the strength of the balloon force. Depending on the sign of β, the balloon

force acts as an erosion or as a dilation. The above geometric curve evolution is efficiently implemented using *level set* methods [145]. The use of level sets has various advantages, with more important that topological changes of the evolving curves (i.e. splitting or merging) are automatically handled. In this theory, the evolving planar curve $\vec{C}(t)$ is represented as a level line of a scalar embedding function $u(x, y, t)$ defined on the whole image domain. The evolution of the contour is done implicitly as we evolve $u(x, y, t)$ under a suitable law. In the specific case of GAC model, the curve evolution (2.10) can be described by the following level set function PDE:

$$\frac{\partial u(x,y,t)}{\partial t} = \left(\mathrm{div}\left(g(\|\nabla I(x,y)\|)\,\frac{\nabla u}{\|\nabla u\|}\right) - g\beta\right)\|\nabla u\| \tag{2.11}$$

This GAC model shows an improved performance compared to Snakes model, as it is parametrization-independent and it can easily handle topological changes (due to level sets). On the other hand, even GAC model has some important limitations, because it considers only local information (based on the simple stopping factor g), without taking into advantage any prior knowledge for the specific type of object we want to detect. In addition, a specific initialization step is still necessary, where the initial curve should lie completely exterior or interior to the object boundaries.

2.4.3.3 Incorporating Prior Information

The segmentation result can be significantly improved by incorporating prior knowledge for the object to be detected, like shape knowledge. Such a prior knowledge can be effectively incorporated in the framework of level sets, as for instance is done in [146], where it is used feature based information and in [147–149] where shape cues are exploited. Consequently, for the specific application of vocal tract image processing, the knowledge about the shape of the vocal tract and of the other anatomical structures in its vicinity can be used from such methods to improve the precision of the corresponding shape recovery.

An example of this is approached of Thimm and Luettin [150] that is based on a Canny edge detector. A set of state images representing images for which the relevant contours have been extracted and checked are used to guide the tracking algorithm. The front teeth, jaw and lips are located and define a background image. The background images is then subtracted from all other images, in order to enhance the tongue contour. As the consistency of contours along time is a decisive help for human experts, a distance between contours in consecutive images has been defined. The tracking thus incorporates two sources of information: the distance between the set of contours detected and a state image and the consistency between the contours obtained at $t-1$ and t. In addition, this temporal consistency is applied in the forward and backward directions in order to be more robust against fast tongue movements, i.e. when the sampling period does not enable any temporal continuity between contours at time t and $t-1$ to be preserved.

A result obtained on one complete film can be seen at:
http://www.idiap.ch/machine_learning.php?project=64.
The evaluation conducted by Thimm and Luettin shows that the contours of lips and tongue have been successfully located in more than 98% of the frames with a sufficient precision.

Fontecave and Berthommier [151, 152] proposed to use a set of key images in which the

tongue was extracted by hand to guide the extraction of unknown images. The guided extraction relies on the Euclidean distance in video features between the unknown and key images. The discrete cosine transforms that represent the tongue contours are a set of 10 points (two coordinates for the two points corresponding to the tongue apex and only the ordinate for the other eight points). The tongue contour is reconstructed by applying a spline to connect the 10 points. Several improvements have been added to reduce the reconstruction error: video and geometrical features have been filtered by a low-pass filter in order to keep only relevant deformations and the indexation is carried out on a neighbourhood of three points to reduce discontinuities.

The approach has been extended to all the speech articulators (lips, velum and tongue) visible on X-ray images. One result is available on
`http://www.icp.inpg.fr/~berthom/m2p/icslp06/tongue-lavals43.html`.
The average error for the tongue is less than 9 pixels, which corresponds approximately to 3 mm. The manual processing of 100 key images (out of the 5673 images of the film) lasts approximately 2 hours.

2.4.4 Morphological Methods

For image enhancement, multiscale analysis and the extraction of geometrical features, techniques from mathematical morphology [153–157] can be used. This is a powerful, novel and effective nonlinear methodology that is based on set and lattice theory and aims at the quantitative description of the geometrical structure (e.g. shape, size) of images and visual objects. Its mathematical background is based on lattice algebra, nonlinear filtering, integrated and stochastic geometry. It offers powerful nonlinear multiscale analysis techniques that have the advantages over linear multiscale approaches of edge preservation and precise size determination. It also provides powerful segmentation schemes based on the watershed transform and similar approaches [158–160]. Mathematical morphology helps in developing object-oriented nonlinear operators for image smoothing and simplification (e.g. the *levelings* [161]), object marker detection for segmentation seeds, watershed segmentation using flooding-type growing based on contrast, size or mixed criteria [162], and region-based shape descriptors. In addition, it combines well with the PDE approaches to smoothing and segmentation problems. For example, multiscale morphology can be modeled and implemented using nonlinear PDEs that are the same or very similar to the ones used for in several nonlinear scale-spaces and geometric curve evolution schemes [163–165]. A promising approach is to further explore the combination of Mathematical Morphology and PDEs used for multiscale analysis and curve evolution. Geodesic curve evolution [143, 166] using level sets [167] can model object shape deformations and detect/track multiple fronts and object-regions in face and vocal tract images and videos by propagating parameter-free active contours. A significant improvement in the boundary detection and segmentation problems could result by combining the attractive features of both approaches into one mixed approach, where the region-based morphology approach offers good image simplification, region markers, and global segmentation whereas the PDE-based contour evolution approach offers better localization and regularity of individual region boundaries.

2.5 Face image feature extraction techniques

We describe next Computer Vision techniques for the automatic extraction of articulatory visual information from *parametric models of shape and texture*. They are applicable to both 2D and 3D modeling applications, but the discussion here is mostly confined to the 2D case.

Parametric models of shape and texture, such as Active Appearance Models [168], Active Blobs [169], Morphable Models [170] and other similar techniques [171–173] are diverse tools for object appearance modeling. Employing a number of parameters controlling shape and texture variation, these models bring a novel image into registration with a reference template, even in cases that the novel image is a deformed version of the template; imaging conditions such as camera position and object illumination can also differ significantly between the template and the novel image. It is notable that these models can represent shape and texture variability in a whole class of objects, such as faces, and learn this variability during a training phase. Such parametric models can be applied to both image synthesis and analysis problems. For example, they have been successfully utilized in applications such as object tracking in video [169], face synthesis [170] and face recognition [174].

An important issue with parametric shape and texture models is to fit them to novel images by minimizing the discrepancy between observed and synthesized appearances. This is a difficult optimization task and general-purpose optimization procedures such as stochastic gradient descent [170] can be inefficient. A rather simple ad-hoc approach that works well in practice is to assume that there is a linear relationship between the error image and the parameter increments; this mapping is learned in a precomputation phase and is used subsequently unaltered, resulting to a very efficient class of algorithms, reviewed.

Parametric models of appearance are generative models which use a compact set of parameters to describe the manifold of shape and texture variation of images depicting a single object or a class of objects.

2.5.1 Object Appearance Modeling

Typically the shape of the object is sampled at L landmarks, whose coordinates constitute a shape vector s of length $2L$ in the two-dimensional case. We allow a particular instance of the shape s to deviate from a mean shape s_0 by letting $s - s_0$ lie in a linear subspace spanned by n eigenshapes s_i, yielding:

$$s = s_0 + \sum_{i=1}^{n} p_i s_i \tag{2.12}$$

The modes of shape variation s_i can be statistically learned using a training set [168], computed by modal analysis of the shape mesh [169], or, finally, selected a-priori to allow for modeling of certain distortions [175]. Often these modes do not model scale and translation, in which cases an explicit similarity transform S_t makes the model scale and translation invariant (S_t has 4 degrees of freedom $t_{1:4}$ with the parameterization of [168, 176]). The enhanced shape parameter vector $\tilde{p} = [t_{1:4}, p_{1:n}]^T$ with length $4 + n$ implicitly defines a dense continuous deformation field $W(x, \tilde{p})$ that maps every point x on the reference object to its corresponding point on a novel object, as follows: The deformation $W(x, \tilde{p})$, which usually is a thin-plate spline or a (piecewise)

affine mapping, is determined by requiring that it maps each landmark in the reference shape s_0 to its corresponding landmark in s. The texture part of the appearance refers to the intensity or color (other information channels can also be added) of the object in a shape-normalized frame. Similarly to shape, allowable texture samples $A(x)$ are generated by a linear model, using a mean texture $A_0(x)$ and a set of m eigentextures $A_i(x)$:

$$A = A_0 + \sum_{i=1}^{m} \lambda_i A_i, \tag{2.13}$$

where we have used vector notation for textures; *e.g*. A_0 denotes the mean texture image raster-scanned into a vector with N entries, as many as the texture samples of the reference object. The eigentexture images, among other things, compensate for illumination changes [171] and model texture variability between different objects of the same class (*e.g*., faces) [168,170]. For example, we show in Figure 2.8 the eigenshapes and eigentextures we obtained by training a model on a person's face. Texture gain and offset can be accounted for separately by a simple texture transformation $T_u(I) = (u_1 + 1)I + u_2$. We gather all texture parameters in an enhanced texture vector $\tilde{\lambda} = [u_{1:2}, \lambda_{1:m}]^T$ with length $2 + m$.

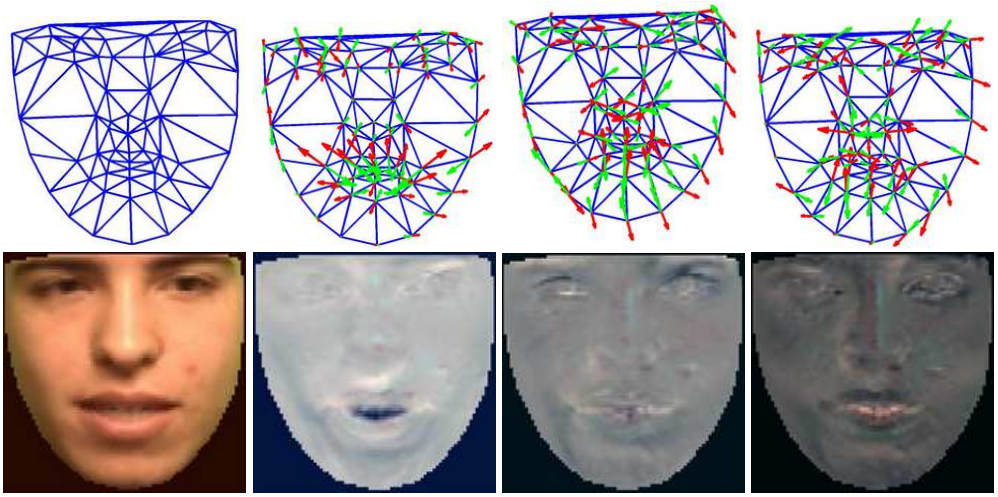


Figure 2.8: *Upper row*: Mean shape s_0 and the first eigenshapes s_i. *Bottom row*: Mean texture A_0 and the first eigentextures A_i.

2.5.2 Model Fitting

A central issue with parametric appearance models is to find algorithms that efficiently and accurately fit them to a novel image I, *i.e*. find the concatenated shape and texture parameters $q = [\tilde{p}^T, \tilde{\lambda}^T]^T$ with length $n + m + 6$ that minimize the discrepancy between the warped-back normalized image texture $T_u(I(W(\tilde{p})))$ and the synthesized texture A. The error image $E(q)$ is:

$$E(q) = T_u(I(W(\tilde{p}))) - (A_0 + \sum_{i=1}^{m} \lambda_i A_i) \tag{2.14}$$

Note that the shape warp $W(\tilde{p})$ denotes the full shape transform $W(\tilde{p}) \equiv S(t) \circ W(p)$, namely deformation followed by similarity (in that order). The mismatch is usually quantified by the Euclidean norm $\|E(q)\|^2$ (sum of square differences) of the error image. Robust norms are necessary when handling occlusion [172]. Minimizing this mismatch is a non-linear least-squares problem on a high-dimensional space and general-purpose optimization techniques such as stochastic gradient descent [170] can be slow. Most efficient techniques to solve the problem require as input a good starting guess for the unknown parameters q and then iteratively update them until a (local) minimum of the mismatch norm is reached. Utilizing an image pyramid and working in a coarse-to-fine strategy increases the robustness of most methods. A popular general technique for improving the parameter estimate is by using a first-order Taylor expansion $E(q + dq) \approx E(q) + \frac{\partial E}{\partial q} dq$ and then applying a Gauss-Newton type algorithm to compute an additive increment by $dq = -K(q)E(q)$, where $K(q) = (\frac{\partial E}{\partial q}^T \frac{\partial E}{\partial q})^{-1} \frac{\partial E}{\partial q}^T$ [177]. However this is computationally very expensive, since image gradients $\frac{\partial I}{\partial x}$ and warp Jacobians $\frac{\partial W}{\partial p}$ need to be recomputed every step [178]. Although $K(q)$ is not constant in general, a number of authors make the assumption that there is a constant linear relationship between dq and $E(q)$ and compute it by multivariate analysis on the training set [168, 169, 175]. Despite its crudeness, this approach leads to very efficient algorithms which often demonstrate good accuracy. However, as Baker and Matthews have noted [176, 178], the so-called *forwards additive* (FA) class of algorithms just described is not the only viable parameter update strategy. They unified previous work on *forwards compositional* (FC) [179] and *inverse additive* (IA) [171] parameter update strategies in iterative image alignment algorithms and introduced the *inverse compositional* (IC) parameter update technique, where a warp parameter update $d\tilde{p}$ is computed to update the warp $W(\tilde{p})$ *compositionally* by:

$$W(x, \tilde{p}) \leftarrow W(x, \tilde{p}) \circ W(x, d\tilde{p}) \equiv W(W^{-1}(x, d\tilde{p}), \tilde{p}) \tag{2.15}$$

They showed that, although the compositional parameter update (2.15) is obviously more costly than the simple additive update $\tilde{p} \leftarrow \tilde{p} + d\tilde{p}$, each full step of the IC algorithm is overall cheaper than in any alternative approach when texture variation is allowed, because it turns out that most of the quantities involved do not change during the fitting procedure and thus can be precomputed, as will be made clear in the sequel. In contrast, the IA method admits efficient implementation only when a restricted class of warps is utilized (including global affine warp) [171] and the FC method is not as efficient as the other two, since image gradients need to be computed every step; see [178] for further details and [99] for an application of the IC approach to 3D morphable model fitting. Baker *et al*. have introduced in a series of papers [176, 180, 181] two algorithms that fall into the inverse compositional framework and are particularly effective. It is notable that the parametric models discussed herein allow to compute not only the parameters that best fit the data, but also the uncertainty in their values. We can employ as uncertainty in the visual features the uncertainty in estimating the parameters of the corresponding non-linear least squares problem [177, ch. 15]; plots of the corresponding uncertainty in localizing the landmarks on the image for two example faces are illustrated in Fig. 2.9. Accompanying the feature values with their corresponding error bars turns out to be particularly important in multi-cue fusion tasks, when one needs to combine the visual cue with other articulatory measurements and uncertain data should be properly discounted.

Figure 2.9: Tracked face shape and feature point uncertainty.

2.5.3 Low-Dimensional Representation of the Mouth

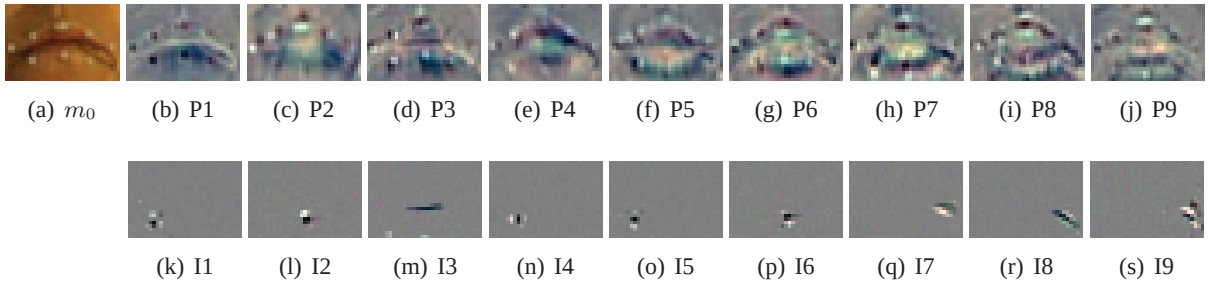
Many methods in audiovisual signal processing rely on extracting the lip contours only [182–186]. The lip contours are modeled using snake-like methods (c.f. section 2.4.3) or data-driven Principal Component Analysis (PCA) methods [182, 184, 185]. An alternative is to not explicitly estimate the shape, but rather the appearance. Saenko et al. [187] employ a cascade of support vector machines that partition lip images according to speaking/non-speaking, closed/narrow/medium/wide, rounded/unrounded, etc. This approach is very robust and enables separation between a small set of spoken commands without the use of acoustic information, but the coarse representation is unsuitable for visual-to-articulatory inversion.

A more viable representation for speech inversion is Independent or Principal Component Analysis (ICA or PCA) of the lip images. The method first stabilizes the image by tracking the head or the lip region, since non-rigid tracking of an articulated face inevitably introduces some errors, due to image noise and necessary simplifications in the model compared to the real face.

The subject's mouth can be stabilized in the images by rigid tracking of the upper part of the face, which usually displays less deformation than the mouth area. Under controlled lighting conditions with the subject facing the camera, a template based 2D method is suitable.

The face pose y_k in frame k can be described by the position, size and orientation of a rectangle over the upper part of the face (above the mouth) in the image. The pattern within the rectangle at frame k, f_k, can be described using a template face pattern f_0 and a probability density function over pose y_k. This density function is estimated in each frame k by iteratively minimizing $||\frac{f_k - f_0}{\sigma_f}||$ using a particle filter [188, 189]. In the first frame of the sequence, the pose value is assumed to be normally distributed over the state-space, which corresponds to initially searching over a wide range of possible poses.

A low-dimensional representation of the image part centered on the mouth can be learned, based on a template image m_0 with neutral lip pose, Fig. 2.10(a). The neutral template is subtracted from each image m_k, with the R, G and B bands subtracted separately. The difference image can be represented as a column vector $x_k = m_k - m_0$ of size d, with $X = [x_1, \ldots, x_N]$, where N is the number of images. A projection of these vectors onto a base $C = [c_1, \ldots, c_n]$, where $n \leq N, n \leq d$ can be expressed as $X \approx CV$ where V is a parameter matrix in the



(a) m_0 (b) P1 (c) P2 (d) P3 (e) P4 (f) P5 (g) P6 (h) P7 (i) P8 (j) P9

(k) I1 (l) I2 (m) I3 (n) I4 (o) I5 (p) I6 (q) I7 (r) I8 (s) I9

Figure 2.10: *(a) Template image. (b-k) The first 9 principal components P1-9. (l-u) The first 10 independent components I1-9.*

subspace defined by C. The base C can be e.g., independent [190] or principal components [182, 184, 185]

Using PCA, C is selected so that the columns represent the n largest principal components (eigenvectors) of the data set, Fig. 2.10(b-j). In ICA, C is instead selected as the n most informative statistically independent components of the dataset, Fig. 2.10(k-s).

Fig. 2.11 shows examples of reconstructions with these two types of component representations for an image of the Qualisys-Movetrack database. In general, ICA is regarded as a better way of representing the lips for visual speech recognition than PCA, for three reasons. Firstly, previous studies have shown ICA to outperform PCA as image representation for visual speech recognition, face recognition and face expression recognition [191–193]. Secondly, the ICA representation is better suited to manage shape difference due to speaker identity [193], since ICA better models the statistical independence of shape differences due to identity and differences due to articulation. Last, the independent component images (Fig. 2.10k-s) are much more spatially concentrated than the principal component images (Fig. 2.10b-j), which means that the representation of the components can be sparsified to speed up the computations. This is essential in a real-time speech recognition system. In the particular case of Fig. 2.11, the PCA reconstruction is however superior, as the lip rounding is better preserved.

Parametric models of shape and texture, as described here, seem to be particularly effective for articulatory visual feature extraction. They have already applied them with good success in facial visual feature extraction for audiovisual speech recognition [85] and automated medical image analysis; see [194] and the references therein for such applications. We strongly believe that this framework can be particularly fruitful for the purposes of ASPI.

(a) Original [ɥ] (b) 100 PC (c) 50 PC (d) 25 PC (e) 100 IC (f) 50 IC (g) 25 IC

Figure 2.11: *(a) Original frame. (b-d) PCA reconstruction of the same frame. (e-h) ICA reconstruction of the same frame.*

Chapter 3

Vocal tract representations

By definition, the output of any acoustic-to-articulatory inversion is a description of the vocal tract shapes. The shapes of the human vocal tract are hopelessly complex and we need a less complicated approximate representation, that is a vocal-tract model. From the birth of speech communication science, researchers have been interested in the relationship between vocal tract shapes and the acoustics, since it constitutes a main part of the observable speech production chain. In order to study such relationships, researchers have needed a simple and thus manageable description of tract shapes. Traditionally tract shapes have been described as area functions or in articulatory models. In the former, the complicated 3D tract is represented by an acoustically equivalent vocal tract area function that is defined by variations of the cross-sectional area along the midline from the glottis to the lip opening. In articulatory models, the vocal tract configuration is described by the state of the articulators, such as the position of the lower jaw and of the tongue. In this section, we shall describe these two types of models.

3.1 Models of vocal tract area functions

The vocal tract area function is defined as the variation of the cross-sectional area along the midline of the vocal tract from the glottis to the lip opening. In this representation, many details such as geometrically complicated cross-sectional shapes and a roughly perpendicular bend between the oral and pharyngeal cavities are neglected. Nevertheless, such a simplified 1D description is acoustically valid for frequencies below 4 kHz, where the main mode of the sound propagations is along the length of the vocal tract. In fact the majority of the vocal tract calculations to obtain speech signals in the time domain and spectral characteristics in the frequency domain employ this simplification with the area function. In a few advanced studies, the vocal tract is described by 2D and 3D geometry in which the main purpose is to gain better understanding of the fricative sound generation and sound propagations of fricative consonants [195], the effect of wall impedance [196], the vocal tract bending [197] or air flow velocity [198].

The description in terms of the area function was formulated to simplify calculations of the vocal tract acoustics. It is common to make further simplification by approximating the

continuously varying cross-section area by a piecewise step function where the length of each section is fixed at 1 cm or less. Since the human vocal tract length is about 17 cm, we have at least 17 variables specifying the area function. Instead of controlling these cross-section variables, it is often desirable to be able to describe the area function with a small number of parameters, which have some articulatory meaning like the tongue position, oral constriction aperture, lip configuration, and so on.

3.1.1 Three-parameter models Xc, Ac, and l/A

About 50 years ago, Stevens & House [199] and Fant [200] proposed a model of area-function having only three parameters, the oral constriction position as the distance Xc from the glottis end, the constriction area Ac, and the length over area of the mouth-opening (l/A). The simplest one proposed by Fant [200] is shown in Fig. 3.1. The total length of the model tract is 16 cm (the larynx tube is neglected) that roughly corresponds to the length of adult male speakers. It can be regarded as a 15 cm long oral and pharyngeal tube with fixed cross-sectional area of 8 cm^2 that is constricted by the tongue-body at the middle section labeled A3, where the area is Ac. Roughly speaking, Xc accounts for the effects of the front/back tongue movements and Ac for those of the tongue height upon the area function. The geometry of the lip section is specified by two variables, the length l and area A. Acoustically speaking however, these two variables can compensate each other and the ratio l/A, which is proportional to the acoustic mass of the opening, is used as the lip parameter. Since the length of the tongue section is fixed at 5 cm, the length of the front and the back cavity is determined from the position Xc.

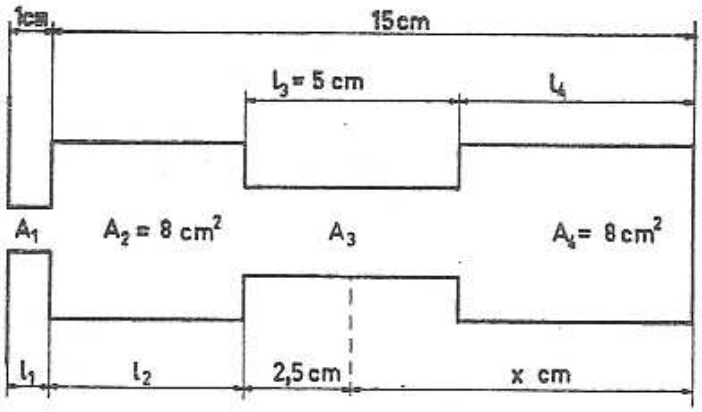


Figure 3.1: *Three parameter model of vocal tract area function proposed by Fant [200].*

Despite of its extreme simplicity, the model captures basic aspects of the articulatory configuration for vowels. In an improved version, the uniform tongue section at the middle is replaced by a smooth parabolic function [199] or by a hyperbolic function [200] to represent more naturally the effect of a rounded tongue-body shape on the vocal tract area function. In these models, the Xc and Ac specify, respectively, the position of the most constricted position and area.

Using this kind of models, Stevens & House [199] and Fant [200] studied mapping from articulatory configurations in terms of Ac, Xc, and l/A to frequencies of the first three to five formants, which were presented as the famous nomograms (for example, in page 76-77 and 82-84 in [200]). These three parameters are sometimes still used to quantitatively characterize the vocal tract geometry of vowels, which are derived from the observation of, for example, X-ray or MRI data, or from a more sophisticated anthropological articulatory model.

Moreover, the nomograms are still referred in the literature when the articulatory-to-acoustic mapping is the issue. It is not unreasonable to suspect that Stevens' quantal nature of speech [201] was inspired by such nomograms. Stevens asserts that the articulatory-to-acoustic mapping is not homogeneous. Instead, in some regions the acoustics is relatively stable against a change in the value of an articulatory parameter and in other regions an abrupt change in the acoustics can occur as the articulatory parameter varies slightly. Stable regions are therefore favoured as target positions for vowels. Recently, phoneticians and phonologists have begun to pay attention to this quantal theory to investigate the origin of sound inventory of different languages. If the theory is correct, vowels situated in such a stable region could manifest one (acoustics) to many (articulatory states) mapping relations, filling up an acoustics-to-articulatory look-up table of an inversion method.

3.1.2 Distinctive region model (DRM)

As described above, the primary motivation of the three parameter models is that they specify the vocal tract configurations with a minimum number of parameters that are interpretable in phonetically relevant articulatory terms, such as tongue position and height, and lip shapes. The Distinctive region model (DRM) was formulated with a completely different philosophy [202]. Mrayati, Carré and Guerin sought the most efficient ways to modulate resonance, and thus formant, frequencies by deforming a given acoustic tube. The formant modulation efficiency is assured by a sensitivity function, which is defined as a ratio of a formant frequency change over a small increase or decrease in cross-sectional area at a point along the length of the tube. In the case of a uniform tube with one end closed (glottal end) and the other end open (mouth opening), the sensitivity function of each formant can be analytically derived and has the form of a cosine function.

Fig. 3.2 depicts sensitivity functions for the first three formants in a binary format considering only the polarity. As the cross-sectional area increases slightly, the formant frequency also increases in plus regions but decreases in minus regions and vice versa. The authors remarked that if the acoustic tube is divided into the eight regions as indicated by the thin vertical lines in Fig. 3.2, an area increase or a decrease would result in distinctive patterns of formant changes. For example, an increase in the region A (a mouth opening) would results in an increase of all the three formant frequencies. Interesting, a decrease in the cross-sectional area in the opposite end of the tube, marked by $\bar{A}$ (at the glottal end), would result in the same formant frequency changes. This is due to the reciprocity property of a close/open acoustic tube, which holds for the pair of regions, such as A and $\bar{A}$. The name of the model, DRM, therefore comes from the fact that a change of the area in each of its regions produces distinctive formant frequency changes.

Exploiting this acoustic property of DRM, the authors formulated a set of strategies to con-

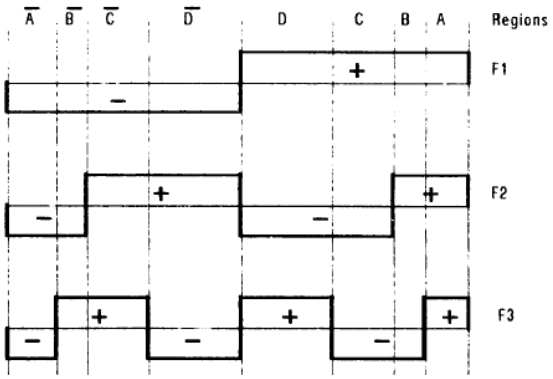


Figure 3.2: *Eight distinctive regions along in a uniform acoustic tube as a neutral vocal tact. The left end corresponds to the glottis and the right end to the lips. A change in the area of each region results in the distinctive formant pattern. (after [203])*

trol formant transitions from one vowel to another [203]. It may seem superfluous to employ eight degrees of freedom to specify the shapes of the vocal tract area functions, but the functional degrees of freedom could be reduced by imposing the control strategies. Although this combination works rather well as a speech production model, the benefit of DRM in the inverse problem is not clear. A study appears to indicate an advantage of DRM in inversion in comparison with a uniformly divided eight-section model of the area function, as discussed latter in Chapter 5.

3.2 Geometrical articulatory models

The vocal tract shapes are determined by maneuvering the articulators such as the lower jaw, tongue, lips and larynx. It is then natural to build a model that can more directly describe the articulatory processes during speech production by human.

As the first such an attempt, Coker & Fujimura [204] formulated an articulatory model describing vocal tract profiles combining simple geometrical elements. It had a circle representing the tongue body that could move in the two-dimensional space delimited by a fixed outer circle representing the palate and rear pharyngeal walls, which is called a ”circle-in-circle” model. The profiles of the lip and larynx were added using combinations of straight lines to complete the articulatory model. In the original version, the circular tongue was specified by the two parameters, front/back position and the height. So, its control parameters were quite similar to the three parameter models of the vocal tract area functions. Its presentation was however closer to the human vocal tract and more natural than the three parameter models of area functions. In fact an elaborate version of this circle-in-circle model, including jaw parameter, was used to calculate vowel-to-vowel formant transitions in a text-to-speech system that employed a formant synthesizer [205]. To calculate the transitions of formant frequencies, the area function

was successively derived from a time-varying configuration of the model.

Mermelstein [206] improved the circle-in-circle model by adding the jaw component and other new features, such as parameters specifying the hyoid bone and the tongue apex position, for studying speech production process. The model described the detailed vocal tract shapes in the midsagittal plane by specifying the value of "a rather abundant set of variables". In practice, many variables were fixed constant values and therefore the actual number of control parameters was often less than 10. The lower jaw is no doubt the most basic component in human articulation system. It influences directly the tongue position, since the tongue sits on the jaw, on lip shapes, and even on the larynx position [207]. In the original articulatory model without the lower jaw as an explicit parameter (e.g., [204]), the effect of the jaw position is implicitly included in the value of each parameter as the tongue front/back position and height, lip tube, and larynx position in a redundant fashion. Since the study of the coordination among different articulators is essential to understand the speech production mechanisms, the explicit use of the jaw position as an independent control parameter is a necessary step forward in articulatory modeling.

These models appear to be capable of describing vocal tract configurations for vowels and consonants with a relatively small number of articulatory parameters. They have weakness in two respects, however. In geometrical models, it is not evident how to specify the parameter values. For example, the parameter values might be determined from vocal tract profiles observed using midsagittal images from X-rays or MRI. The determination would require, however, a fitting procedure, manual or automatic, between model and data [208]. The parameter values cannot be calculated from data in a straight forward way. Moreover, the models were built in an ad-hoc manner based on the authors' knowledge and good intuitions. It is therefore not so simple to evaluate the adequacy of these models. These weaknesses make another approach interesting, that of modeling based on analysis of articulatory data, which we shall discuss in the following section.

3.3 Data-based articulatory models

In general, articulatory data, for example tongue profiles, are composed of a mass of apparently unlawful curves. Such data however often contains a lot of redundancy. It is useful therefore to perform a data reduction in the analysis so that the data are described by a set of orthogonal or uncorrelated variables. If those variables were interpreted in articulatory terms, the concise way of describing the raw data could be regarded as an articulatory model and the variables as articulatory parameters.

3.3.1 Models based on Fourier Coefficients

Heinz and Stevens [209] were the first to analyze X-ray data with a semi-polar coordinate system, i.e. consisting of a polar part covering the oral cavity and a Cartesian part for the pharyngeal region. The shape of the tongue profiles from the tip to the root was aligned with respect to fixed anatomical landmarks visible in the X-ray pictures, such as the upper incisors. The coordinate system was then placed on each of the aligned profiles and the shape was measured

as a series of intersection points between the tongue contour and the coordinate grids. Each tongue shape can therefore be represented by a vector that consists of the radius in the polar system in the mouth region and the distance from the vertical axis of the Cartesian system in the pharyngeal region. The sampled original contours can be recovered by the projection of the corresponding vectors onto the semi-polar coordinates. The use of such a coordinate system for measuring shapes on image data obtained using X-rays, MRI, and ultrasound has now become a standard procedure.

Liljencrants [210] investigated static X-ray pictures during the production of 10 Swedish vowels by two subjects and observed that tongue vectors plotted in function of the element (coordinate) numbers were quite smooth, indicating that each vector is mainly composed low spatial frequency components. This observation motivated his Fourier (coefficient) representation of the tongue vectors. To recreate the original tongue contours, the tongue vector is resynthesized from the sine/cosine coefficients (equivalent to magnitude and phase) and then projected on the semi-polar coordinate system. If N/2-1 pairs of coefficients were used (where N is the size of the tongue vector), the recovery is error free. If some errors in the re-synthesis are acceptable, higher-order coefficient pairs may be truncated. In fact, Liljencrants' experiment showed that one pair of the cosine and sine coefficients at the fundamental spatial frequency can already describe the observed tongue vectors with an acceptable root mean square error. With the two lowest coefficient pairs, the error is negligible.

The Fourier model hence provides a mathematically simple and elegant solution to describe tongue contours with high efficiency, but the sine and cosine functions are difficult to interpret in articulatory terms. As described in the following, factor analysis approaches provides an articulatorily interpretable model with the elegance of Fourier model.

3.3.2 Models based on factor analysis

Harshman, Ladefoged, & Goldstein [211] analyzed tongue profiles of English vowels using a procedure PARAFAC based on the principal component analysis (PCA). It identified two major factors. As shown in Fig. 3.3, one factor accounts for a forward movement of the back of the tongue concomitant with an upward fronting movement of the tongue blade. The second factor accounts for a forward movement of the tongue root associated with the upward backing movement of the tongue body. The directions of these two movements appear perpendicular to each other, which is a consequence of the imposed orthogonal (or uncorrelated) nature between factors. The observed tongue profiles are described by the weighted sum of these two components. In this study however, the factor that accounts for the influence of the lower jaw movement were not extracted from the tongue profile data.

Kiritani, Sekimoto, & Imagawa [212] explicitly treated the jaw factor in the analysis of the X-ray microbeam data for Japanese vowels. In this study, the tongue deformations during vowel sequences were measured by tracking four pellets glued to the tongue surface, i.e., a flesh point measurements. The jaw and lip movements were also measured by tracking a pellet fixed, respectively, on the lower incisors and on the lower lip. All the six pellets were aligned in the midsagittal plane. Although the observation of jaw, tongue, and lip movements with these six pellets appears under-sampled, this study has shown how a jaw-based articulatory model could be formulated by a statistical analysis on articulatory data.

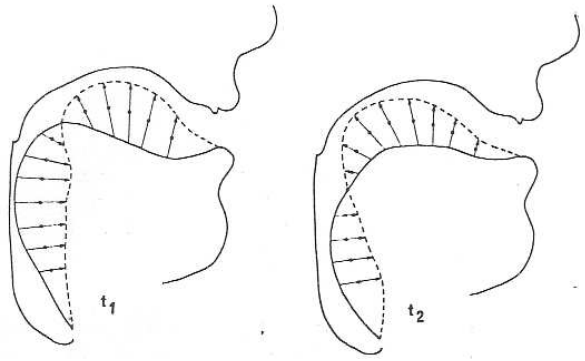


Figure 3.3: *The effects of the first two principal factors upon the tongue profiles. The tracings show the tongue position for a large positive (solid line) and a large negative (dashed line) factor value. The dots in the arrows from positive to negative shape corresponds to the zero factor value, i.e. an average (neutral) tongue contour. (after [211])*

In the analysis, the measured jaw position in terms of x- and y-coordinates was first approximately described by a single number, i.e., jaw parameter J, using its projection on the regression line that was calculated on the entire jaw position data of a given speaker. Second, linear regression lines are determined between the jaw parameter and the x-coordinate value (and y-coordinate value) of each of the six pellet positions, including the jaw pellet. The component explained by each linear regression represents the jaw-dependent displacement on the x- or y-coordinate of the observed pellet position. Third, these displacements are subtracted from the corresponding six pellet positions. Finally, the residual displacements were analyzed by the PCA. Fig. 3.4 illustrates the individual effects of the jaw factor J (determined by the regression analysis) and of two principal factors, T1 and T2, upon those 4 tongue pellets. It is interesting to note that the first principal factor T1 accounts for the front/back tongue-body movements and the second factor T2 for deformations, bulging vs. flattening, of the tongue body. The effects of the jaw parameter J and of T2 are somewhat similar In detail however, J affects the height of the tongue, whereas T2 the front/back position.

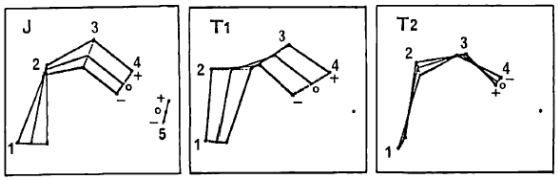


Figure 3.4: *The effects of the first three factors upon the tongue profiles. (after [212])*

The main goal of those statistical analyses was to study speech production processes with simpler and manageable representation than the raw articulatory data. Tongue deformation during the production of vowels can thus be investigated by two principal components [211] or by the jaw parameter and two principal factors [212]. These statistical models can be regarded

as articulatory models as long as the effect of each factor on the tongue contours can be interpreted in articulatory terms.

3.3.3 Vocal tract models based on X-ray data

Maeda [213] created a complete vocal tract model based on the factor analysis of X-ray data. First for a set of isolated vowels in earlier versions and in later versions for a set of 10 short natural French sentences [214]. The later versions actually consist of three models, for the tongue, lips and larynx [215]. Since these three articulators can be assumed to be independent, except that they are affected by the lower jaw position as mentioned before, they are analyzed separately with the jaw position as a common factor. Here we cannot use classical PCA, since it does not allow us to specify the measured jaw position as a factor. Maeda [213] employed an arbitrary orthogonal factor analysis proposed by Overall [216], which is now refered to as guided PCA. It operates on the correlation matrix derived from observations of variables, such as the tongue vectors. The matrix is assumed to be the linear sum of correlation structures and each structure is determined by a factor pattern that accounts for the first-order effect of a cause, such as the jaw position, upon the observed variables. In this view, the PCA determines the structures so that the maximum of variance is explained by each factor.

Let us explain how this analysis works with an example. Fig. 3.5 shows a typical method to measure the shape of tongue contours applying the semi-polar coordinate system, which is fixed relative to the head position. A set of the value of intersection points between the coordinate grids and the contour (which is a vector) represents the contour shape. The figure also shows how the jaw position is defined. The line connecting the upper and lower incisor tips is projected to the straight line having an angle theta. This projection J is considered to be the lower jaw position (or opening). The vocal tract state is represented by a vector that consists of the tongue coordinates, J and the exterior tract walls. This kind of vectors is collected for all frames in the X-ray data.

In this particular study, the angle theta was determined to maximize the influence of the J parameter on the variance of tongue vectors, which turned out to be 65°. Next the correlations among all the variables, i.e., tongue vector elements plus J, are calculated to obtain a correlation matrix. The correlations between the J and the tongue variables are then subtracted from this original correlation matrix. This subtraction assures that the subsequent factors are uncorrelated [216]. The influence of the J opening on the tongue contour is depicted in Fig. 3.7a. J explains nearly 30% of the variance, indicating an important contribution of the lower jaw upon observed tongue shapes.

Fig. 3.6 shows how the other factors influence the variance, after the extraction of the effect of the jaw position J. Two options are now available to identify tongue factors, standard [215] or guided PCA [213]. Standard PCA results in a greater value of the variance explained by the first three PCA components than the guided PCA. These three components, i.e., PCA determined intrinsic tongue factors, could be interpreted, in order, as tongue front/back, tongue-dorsum shape, and the tongue apex, based on their influence on the tongue contours.

For the arbitrary orthogonal factor analysis, the variance in the tongue shape instead determines which variable that is chosen manually. In order to select the best tongue variable,

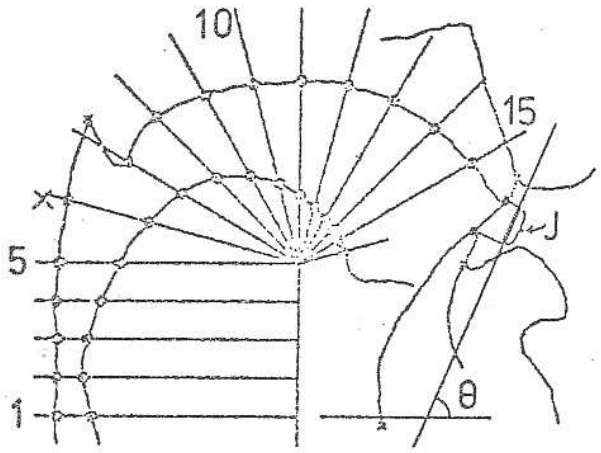


Figure 3.5: *Measurement of the form of vocal tract inner and outer contours using the semi-polar coordinates and of the jaw parameter J.*

the variance explained by each tongue variable, from the coordinate number from 1 to 15 (see Fig. 3.5 for these numbers), is calculated. The result is shown by the curve connecting circles in Fig. 3.6. The curve shows that the tongue variable at the coordinate 5 extracts the highest variance, about 38%. The tongue variable at coordinate 5, therefore, is the second factor and its influence is shown in Fig. 3.7b. Since the value at the coordinate 5 should be a reasonable measure for the front/back position of the tongue, this factor is called the [front/back] factor.

Now, the correlation structure explained by this front/back factor is subtracted from the residual correlation matrix, resulting in a residual variance shown by the curves connected by filled circles in Fig. 3.6. Note that the value at the coordinate 5 is absent, since its variance was exhausted by the previous jaw and the front/back factors. The calculated variances show the maximum value at the coordinate 9, which explains about 19% of the variance. Fig. 3.7c indicates its effect on the tongue contour. This third factor appears to control the tongue-dorsal shape, and can be called the tongue-dorsal factor. This factor is important for the production of high-back vowels, such as /u/. The correlations explained by this third factor are subtracted from the residual correlation matrix again.

As seen in the curves connected by the triangles in Fig. 3.6, the next factor (at coordinate 14) extracts about 9% of the variance. Fig. Fig:Maeda3d shows its effect on mainly the tongue apex, and is therefore called the tongue-apex factor.

Note that these three intrinsic tongue factors explain 38+19+9=66% of the variance and with the Jaw factor 96%. It is hence reasonable to stop the extraction procedure after the fourth factor. Moreover, it can be stated that the jaw-based tongue contours can be best predicted using the values of those three coordinates, coordinate 5 in the back of the tongue, coordinate 9 in the palatovelar region, and the coordinate 14 near the apex.

For the sake of simplicity, the lip opening is described by three variables: the minimum separation (distance) of the upper and lower lip (a lip tube height), the distance between the front

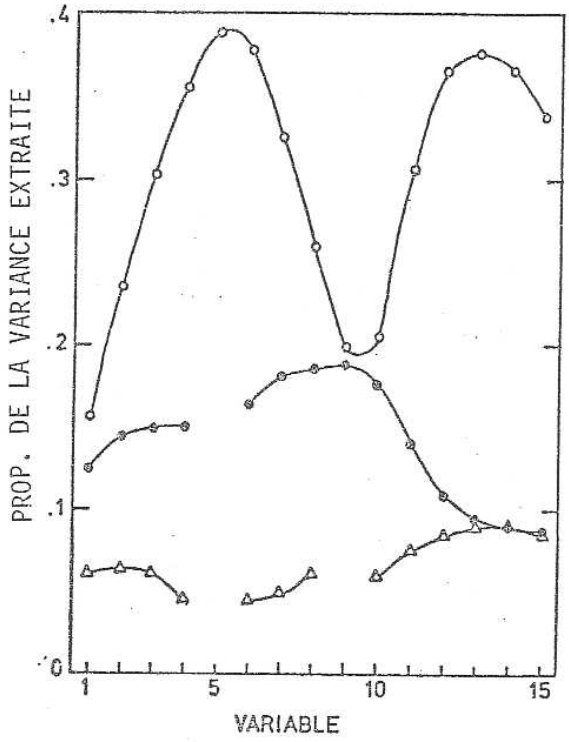


Figure 3.6: *Proportion of variances explained by assuming 15 tongue variables as factors (parameters): tongue-body position in the open circles, tongue-body shape in the closed circles, and tongue apex position in the triangles.*

upper incisor and the minimum separation point (a lip tube length or lip protrusion), and the lip tube width measured on frontal photographic images of the subject face. These variables are predicted by the extrinsic jaw parameter and two intrinsic lip parameters, the height and protrusion. Only the lip-tube width, which is not visible in the midsagittal X-ray data are predicted by the model derived by the factor analysis. In addition, the front and the back edges of the glottis described by x- and y-coordinate values relative to the head are used to define the position of the larynx and the one end of the vocal tract. The four variables are predicted by the extrinsic jaw parameter and the intrinsic larynx height parameter. The derived articulatory model of the complete vocal tract is shown in Fig. 3.8.

The output of the inversion in the form of an articulatory model, geometrical or statistical, can be evaluated against articulatory data, if such was recorded together with the acoustics that was the input to the inversion. However, if articulatory data does not exist for the sequence, the evaluation has either to be based on general phonetic knowledge about the plausibility of the estimated articulation, or on an analysis of the resulting acoustics, created through a resynthesis that uses the articulatory model as input.

Speech synthesis is an important component to validate acoustic-to-articulatory inversion, since a time domain acoustic simulation using a vocal tract model allows for acoustic comparisons between synthesized and the speaker's actual acoustics. Experiences in vocal tract syn-

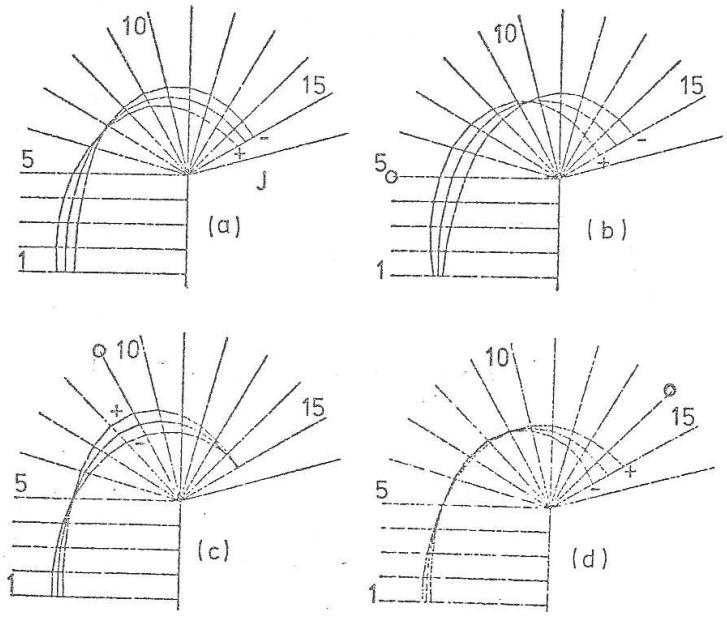


Figure 3.7: *Effects of the four factors upon tongue contours: Jaw factor in (A), tongue-body position in (b), tongue-body shape in (c), and tongue-apex position in (d). In each plate, the symbols '+' and '-' indicate, respectively, the value of +1 and -1 in standardized units. The contour without the symbol represents the averaged (or neutral) tongue contour.*

thesis chould also help us formulate articulatory constraints in the inverse procedures. Since the two-dimensional models describe only the vocal tract shapes in the midsagittal plane, it is necessary to derive the corresponding area function for the resynthesis. In the next section, this midsagittal-to-area conversion is be described.

3.4 Midsagittal-to-area conversion

In theory, the vocal tract acoustics, the waveform in the time domain and transfer ratio in the frequency domain, can be calculated from its three-dimensional (3D) geometric description. So far, the 3D calculations are not so successful, presumably due to the difficulty of specifying an appropriate meshing of the complex 3D vocal tract for the application of a finite element method. The 3D calculations suffer from excessively long computation times, often hours of calculation for a 10 ms long speech signal. For these reasons, the conventional one-dimensional calculations with vocal tract area functions appears to still be the best alternative. Since the major mode of the acoustic wave propagation is along the length of the vocal tact and higher modes, such as transversal propagation, can be neglected in frequencies below 4 kHz and the area fucntions calculations are hence valid.

With a vocal tract configuration described in the midsagittal plane, it is necessary to apply a midsagittal-to-area conversion to derive the area function. The conversion uses the distance

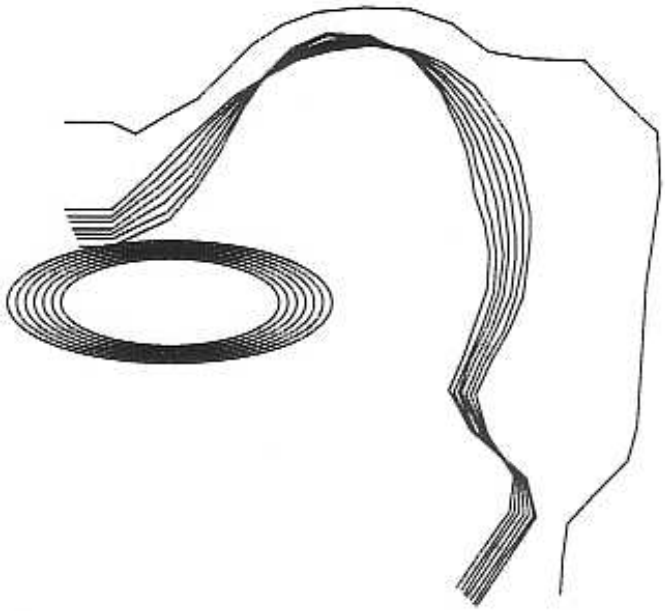


Figure 3.8: *Complete vocal tract profiles derived from an X-ray data derived articulatory model, showing the effects of the jaw parameter.*

along each coordinate gridline between the inner and outer contours. The inner contour consists of the front laryngeal wall, tongue and lower lip, and the outer contour, corresponding to the rear laryngeal and pharyngeal walls, soft and hard palate, and upper lip. These measured sagittal distances (or heights) d (cm) along the vocal tract from the glottis to the lips are converted into areas A (cm^2) assuming a power function [209]:

$$A = \alpha d^{\beta} \tag{3.1}$$

where α and β are parameters that vary depending on the position in a speaker dependent fashion. If the vocal tract had a circular shape, the value of α would be $\pi/4$ and β=2. Since the cross-sectional shape of the vocal tract is complicated and far from a circle, the values of α and β must be empirically determined. Moreover, the actual cross sectional area can depart from the power function scheme as the value of d becomes large [217, 218]. Acoustically this is not so critical however, because the formant frequencies become relatively insensitive to local variations as the area becomes large.

The area function is defined by the accumulated distance between the centre point if the vocal tract air passage at each gridline and the cross-sectional area at that gridline. If necessary, the area function can be simplified by resampling the length (x-) axis at equally spaced sample points. After such a simplification, the area function corresponds to a set of equal length uniform tubes connected, for example circular, end-to-end. The uniform section length is often chosen as 1 cm or less. This apparently gross approximation of the real vocal tract

is valid in frequencies below 4 kHz where the effects of the tract bend and the shape of the cross-sectional area can be neglected.

Even if the area-function representation is acoustically valid at low frequencies, some improvement is possible and sometimes necessary. In the above derivation of area function, the implicit assumption is that the wave front of the acoustic propagations matches the semi-polar grid lines, which is no guaranteed. In the case of a tube with a smooth 90° bend, the wave front can be assumed to be perpendicular to the geometrical midline of the tube. Then it is reasonable to consider that the wave front in the vocal tract is also perpendicular to the midline. Cross-sectional areas determined along the coordinate grids can hence be corrected by the cosine of the angle between each grid and the wave front at that grid.

From the above discussions, it might become clear that the use of the semi-polar coordinates for deriving area function is motivated by the matter of convenience rather than by acoustic principles. The vocal tract midline and the height dimension, d, can instead be determined by assuming a spherical wave front, i.e. a circular front in the midsagittal plane [219, 220]. In this approach, a heuristic algorithm determines a series of circles that just fit between the inner and outer contours of a vocal tract profile. The midline is obtained by connecting the center of the circles along the tract length from the glottis to the lip opening. The height dimension is determined either as the line between the two contact points on the circle with the inner and outer contours or as the line segment passing the circle's center and perpendicular to the midline. This method does not use a coordinate system and the determined tract length tends to be slightly shorter than that determined using a coordinate system. This seems to result in a better match between calculated and measured formant frequencies, although this has to be confirmed on a larger body of data.

Regardless of a method used, the sagittal-to-area conversion involves certain degrees of incertitude. Then one might wonder why an articulatory model can produce vowels with a high phonetic value from a converted area function. This can happen, we think, because errors in the conversion can be compensated, at least in part, by adjusting the values of articulatory parameters. Further we suspect that human speakers must do similar adjustments in articulation in order to compensate for individual differences in the vocal tract morphology. If this is the case, it is not so unreasonable to use this imperfect sagittal-to-area conversion. In the inverse problem, if the purpose is to recover the exact geometry of a subject tract configuration from audiovisual speech however, it is necessary to devise an accurate conversion scheme for each subject.

3.5 Time-varying areafunctions in vowel-consonant sequences

Up to this point, the description has focused on the representation of static sounds. In running speech, however, the articulators are moved asynchronously [221] when going from one phoneme to another, causing the vocal tract cross-sections to vary in a specific spatiotemporal organization. The areafunction is then instead specified by the cross-sectional area $A(k,t)$ and the length $x(k,t)$ at the kth section at time t. In a phoneme sequence, the area function will vary smoothly between the target area functions for the phonemes in the sequence.

For a vowel-consonant-vowel sequence, the target area functions can be represented as

in Fig. 3.9. The target area function for consonants is a uniform tube (corresponding to a neutral vowel) having a constriction formed with a single section kth constricted with a cross-sectional area (A_c) close to zero [222]. The transition between the target area for section k is schematically shown by the thick line in Fig. 3.10. The target area of the initial vowel (V1) is specified at the onset (t_0) and offset point (t_1), indicated by the open circles. During the interval $t_0 - t_1$, the section area is kept at the target value for vowel V1. The target of the consonant is specified at its onset (t2) and the time varying area function in the transition from the V1 offset to the C onset is obtained by interpolating between the two targets with a cosine law. The stationary consonant configuration remains up to its offset at t3 and then smoothly changes with a cosine law to the V2 target at t4.

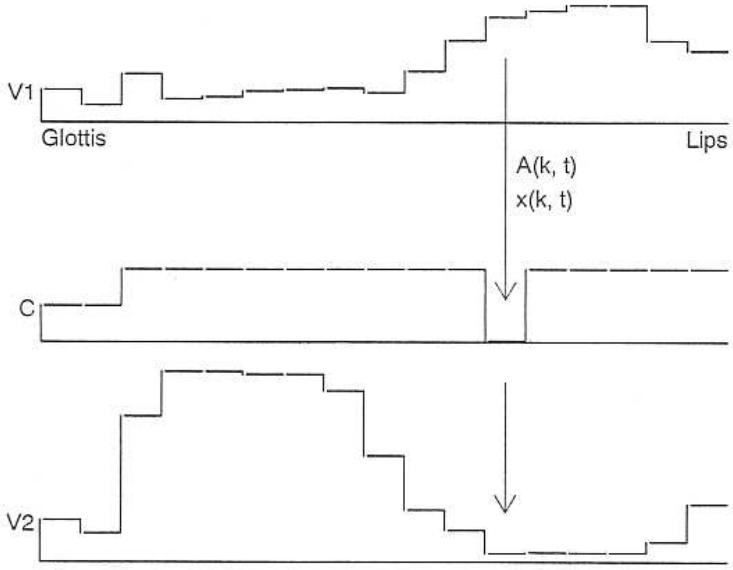


Figure 3.9: *Target area functions in a vowel-consonant-vowel sequence, with a consonant constriction at section k.*

Fig. 3.10 shows the area variation at section k for a fricative and an unvoiced stop consonant. The stop consonant is created with a slight modification of the fricative pattern. The temporal pattern for the stops consists of closure and release frication, specified by two successive target area functions. The temporal variation of the section corresponding to the stop is shown by the dotted line. The total consonant duration is slightly longer than that of fricatives and this lengthening is achieved by shortening of the transitions between vowel and consonant.

Just as consonant articulations are simplified, as described above, it would be possible to use simplified area functions for vowels, such as two or four uniform tubes having different diameters connected end to end. Although isolated vowels synthesized using these simplified area function sound quite correct, the use as targets often resulted in a poorer sound quality than that of realistic vowel target area functions such as those shown in Fig. 3.9. We suspect that the simplified area functions having strong spatial discontinuities produce unrealistic configurations and transitions when they are interpolated section by section. For this reason,

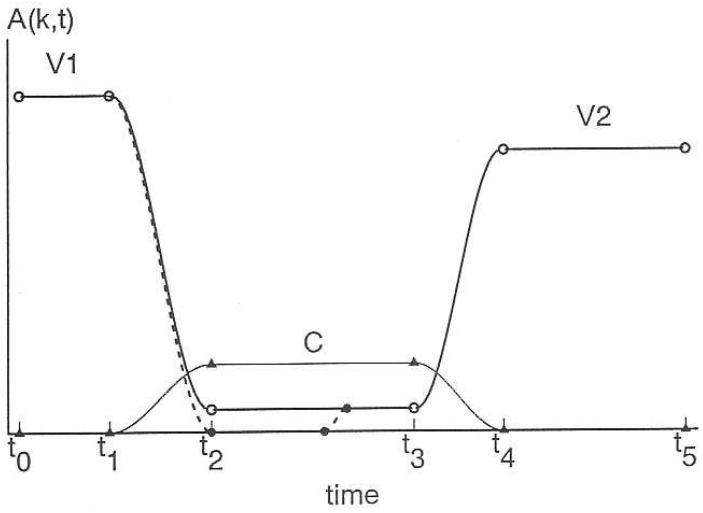


Figure 3.10: *Schematic representation of oral constriction area variations specified by the targets at the "turning" points indicated by the markers. The variation of constriction section in V1-fricative-V2 is indicated by the thick solid line with open circles. The dotted line with closed circles corresponds to V1-stop-V2. The thin line with filled triangles indicates the variation of a slow time-varying component of the glottis, A_{g0}, which is common for the two classes of consonants.*

smooth area functions are preferable for vowel targets.

This area functions interpolation is an important advantage of a vocal tract synthesizer, as compared to, e.g., a formant synthesizer, where vowels are characterized by only poles of the transfer function, whereas consonants require poles and zeros [200] [223]. The interpolation in the formant domain, thus, would become complicated, if one wishes to interpolate the transitions between a consonant and a vowel by properly handling the appearance and disapperance of poles and zeros. With time-dependent area functions, the vocal tract is nothing but a smoothly time-varying acoustic tube that efficiently handles transitions between vowels and consonants. The creation of fricative noise and stop bursts at the glottal and supra-glottal constriction however needs special attention. The computational complexity makes an honest aerodynamic simulation of airflow turbulence impractical for speech synthesis purpose and a functional model of the noise generation is therefore often used, as described in the next section.

3.6 Synthesis of fricatives

The functional model of noise generation is a band-pass filtered sequence of random numbers, which is injected at the exit of the constriction or at one section downstream, depending on the consonant. In the actual simulation, the noise injection is treated as the insertion of a dipole noise pressure source in series [224], which gives a typical short-term noise-source spectrum

as illustrated in Fig. 3.11. The original flat spectrum is shaped by a third-order highpass filter and a first-order lowpass filter. The magnitude of noise must be modulated by the aerodynamic condition.

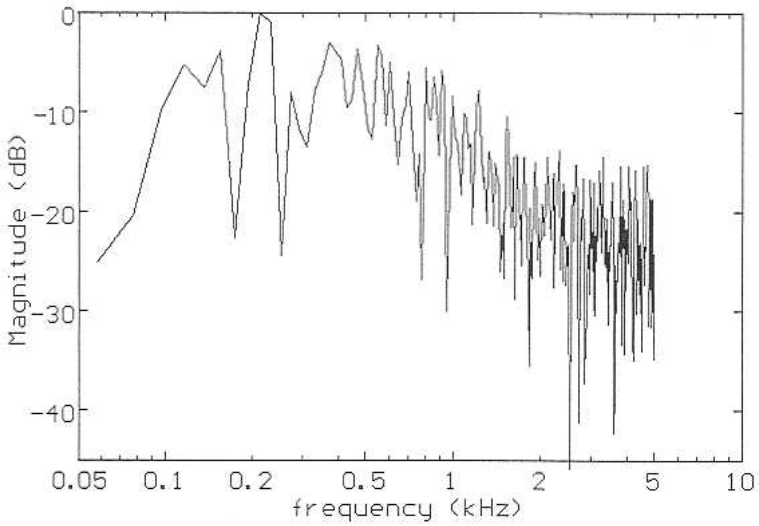


Figure 3.11: *A typical noise source spectrum.*

In practice, the noise magnitude is determined by multiplying a weight (or a gain) which varies as a function of the cross sectional area of the constriction and the airflow level. Either a square law [223] or a cubic law [225] [226] can be used. The simpler square law is used in this study. The gain, N_{mag}, is determined by the following relations:

$$N_{mag} \propto \frac{U_d c^2}{A} \quad or \propto R_e^2 \tag{3.2}$$

where $U_d c$ is a low frequency airflow in cm^3/s, and A is the cross-sectional area of the constriction in cm^2. R_e denotes the Reynolds number. Since the aerodynamic law specifies only proportionality, the value of the scaling coefficient must be empirically determined.

Now in order to calculate the noise magnitude, it is necessary to have the value of the airflow, $U_d c$, inside the vocal tract. It can be calculated using a low frequency model [227, 228], where air flow is determined by the function of the sub-glottal air pressure, P_s and the cross-sectional areas of the two major constrictions, one at the glottis (A_g) and the other in the supra-glottal tract (A_c). The flow resistance at these constrictions can be approximated by the sum of the Bernoulli kinetic resistance, R_b, and viscous resistance R_v. R_b is calculated as

$R_b = (k_b \rho / A^2) U_d c,$

where ρ is the density of air. The scaling coefficient, k_b, depends on the cross-sectional shape, being close to one for a circular duct and 1.42 for a rectangular [229]. The different formulations of the viscous resistance per unit length are needed for the glottis and for the supra-glottal constriction as:

$R_v = 12l_g^2\mu/A_g^3$ (for the glottis)
$R_v = 8\pi\mu/A_c^2$ (for the supra-glottal constriction),

where l_g is the length of the vocal folds and μ indicates the viscosity coefficient of the air. Note that the Bernoulli resistance is a non-linear element because of its dependency on the airflow level. U_dc is therefore obtained by solving the following second-order equation:

$$k_b(\frac{1}{A_g^2} + \frac{1}{A_c^2})U_{dc}^2 + (\frac{12l_g^2x_g}{A_g^3} + \frac{8\pi\mu x_c}{A_c^2}U_{dc}) - P_s,$$

where x_g and x_c are, respectively, the thickness of the vocal folds (i.e., the length of the glottal section) and the length of the constriction section.

As a consequence, the noise magnitude can be automatically modulated depending only on geometric variables of the vocal tract, as section length (x_g and x_c) and section area (A_g and A_c) and on the subglottal air pressure, P_s. Only the value of the scaling coefficient for N_{mag} remains to be determined. The value is empirically determined so that the level of the synthesized frication noise relative to that of surrounding vowels is realistic.

3.6.1 Time variable glottal section

The glottal section is considered as a part of the vocal tract area function, but its time-varying characteristics are quite different from the other sections of the vocal tract. The temporal interpolation of targets, therefore, is differently treated for the area function and for the glottal section.

The glottal section consists of slow and fast time-varying components. The muscular adjustments in the laryngeal system determine the slow adduction/abduction during consonant production. When certain aerodynamic and biomechanical conditions are met, the vocal folds vibrate, which is described by the fast pulsating oscillation of the glottal section in the simulation. The area of the glottal section, A_g, therefore, is specified by the sum of the slow time-varying area, A_{g0}, and the fast time-varying area, A_{gp}.

The correct adjustment of A_{g0} is important for the generation of the fricative noise. Roughly speaking, the airflow U_{dc} specifying noise magnitude remains equal along the entire vocal tract including the glottis for a given instant. Since the noise magnitude is inversely proportional to the cross-sectional area, A_{g0} must be larger than the supra-glottal constriction area A_c for the frication noise to dominate over the aspiration noise. The temporal pattern of A_{g0} is determined by specifying target values at onset and offset points. A typical example is shown by the thin line in Fig. 3.10. During the stationary part of vowels, the value of A_{g0} is kept at zero for the synthesis with a male voice. The A_{g0} can be adjusted to a non-zero value to mimic, for example, a breathy quality of a female voice.

Vocal fold vibration can be simulated using a physical model, such as the two-mass model of Ishizaka and Flanagan [229], but for the sake of computational simplicity, a descriptive glottal pulse model is often used instead. Fant [230] originally described the glottal flow pulse in volume velocity, but for area function modelling, it may be more convenient to use glottis area variations. The pulse shapes in flow and in area are similar to each other, except that the flow has a more skewed pulse shape due to the inertia of air mass [223]. The skewness is important for the spectral characteristics of the voice source and it can be arbitrarily manipulated by

adjusting the opening and closing quotients of glottal pulses. The pulse shape is determined by three variables: the peak amplitude, A_p (cm^2), duration of glottis opening, t_1 (s) and of closing duration, t_2 (s). It is convenient to specify these durations by the quotients of the pitch (fundamental) period, T_0 (s). In a gross approximation, fixed value of quotients can be used, sucg as $t_1 = 0.36T_0$ and $t_2 = 0.26T_0$, which seem appropriate for a male voice.

Using above defined variables, the glottal area-pulse within a pitch period is described as

$A_{gp}(t) = \frac{A_p}{2}(1_c os(at))$ for $0 \leq t < t_1$ (opening quotient),
$A_{gp}(t) = A_p(1 - b + bcos(a(t - t_1)))$ for $t_1 \leq t < t_1 + t_2$ (closing quotient),
$A_{gp}(t) = 0$ for $t_1 + t_2 \leq t < T_0$ (closed quotient),

where the coefficient $a = \pi/t_1$ and $b = 1/(-cos(\pi t_2/t_1))$. During the opening quotient, the glottal area increases smoothly without any discontinuity, which is described by the raised cosine function. The glottis closes abruptly at the end of the closing quotient that corresponds to the main excitation of the vocal tract. Since T_0 is the inverse of the fundamental frequency, F_0 (Hz), the glottal pulse is determined by the function of only two variables, A_p and F_0. In the synthesis, the target values of these two variables at appropriate turning points are specified, and their values at any given time are then calculated by linear interpolations.

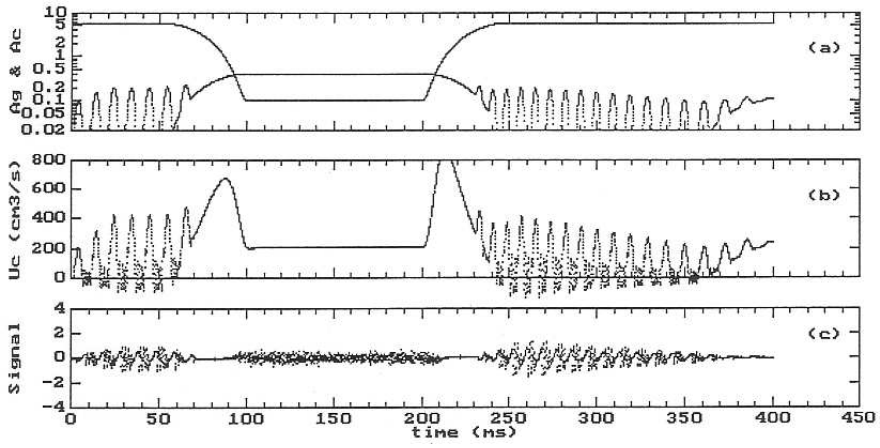


Figure 3.12: *Calculated airflow and speech signal for /asa/. Specified variations of the glottal area, A_g (cm^2), and the constriction area in the alveolar region, A_c (cm^2), are shown in (a), the calculated airflow (cm^3/s) at the exit of the constriction in (b), and the radiated sound in arbitrary units in (c).*

Fig. 3.12 illustrates the result of the synthesis of the sequence /asa/, as an example. A uniform tube having the constriction section at 1 cm from the lips is used as the target for the fricative /s/. Fig. 3.12a shows the temporal variations of the constriction section area (A_c) and the glottal area (A_g) specified by target interpolations for /asa/. It should be noted that the waveform is deformed because the y-axis is logarithmic to cover the large range of area variations. The corresponding simulated airflow at the exit of constriction (U_c) and the radiated sound are shown, respectively, in Fig. 3.12b and in Fig. 3.12c. During vowels, the glottal pulses excite the vocal tract. In theory, noise is generated at the glottis during the open

quotient of each voice period. Apparently this pitch-synchronized noise is weak and not visible on the synthesized wave form in Fig. 3.12c. Toward /s/, the glottis opens up and A_c closes down, thus A_c becomes smaller than A_g during the fricative. Consequently, frication noise is generated at the exit of the constriction. It is interesting to note that two airflow peaks appear during vowel-fricative-vowel transitions. This kind of a double peak is often observed in the natural production of /s/ as shown in Fig. 3.13 The peaks coincide with the crossover points between A_g and A_c. Just before the first crossover and just after the second crossover, the aspiration noise at the glottis is dominant, thus the corresponding radiated noise has different characteristics in comparison with the frication noise. This is seen in the spectrogram shown in Fig. 3.14b as two distinctive noise bands which appear at transitions between vowel and fricative. The spectrogram of natural token also exhibits such characteristic noise bands, as shown in Fig. 3.14a, although the appearance is quite different from the synthetic version. The synthetic token could be made closer to the natural one by adjusting the noise-source spectral shape and the gain coefficient, N_{mag}.

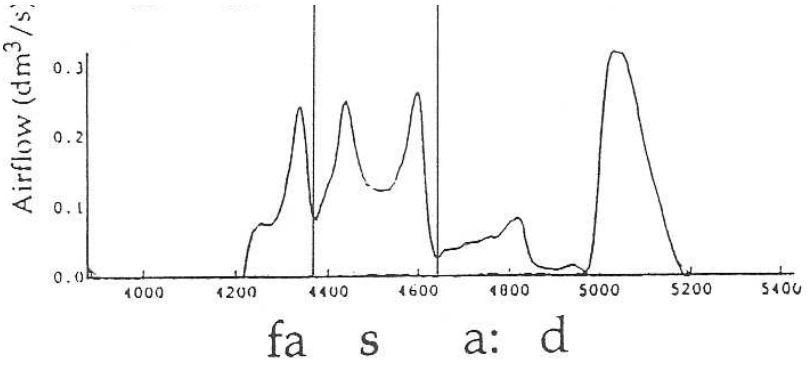


Figure 3.13: *Measured airflow during /fasa:d/ in Arabic, exhibiting a double peak around /s/ (after Yeou and Maeda [231].*

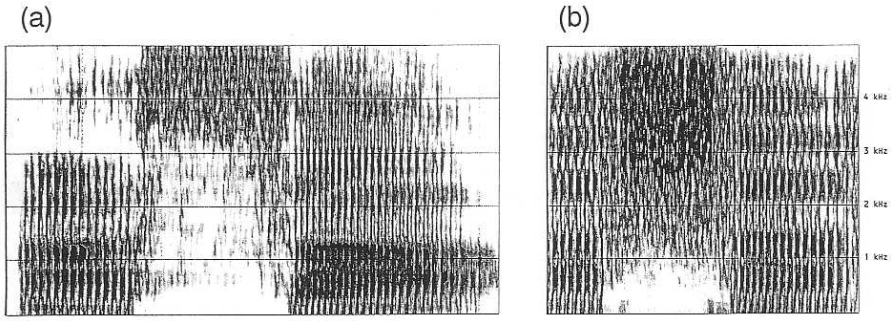


Figure 3.14: *Spectrograms of /asa/: (a) natural token, and (b) synthetic token which corresponds to that shown in Figure 3.12.*

Chapter 4

Acoustic representations

4.1 Representation of spectral parameters

Acoustical simulations of the vocal tract show that the first four formant frequencies are roughly linearly distributed. For an average male speaker F1 is between 300Hz and 750Hz, F2 between 800 and 2000Hz, F3 between 2000 and 3000Hz and F4 between 3500Hz and 4000Hz. Therefore, spectral analysis used for acoustic-to-articulatory inversion should present a good frequency resolution up to 4000Hz.

4.1.1 Linear prediction of speech

The idea behind linear prediction to exploit the correlation between consecutive speech samples to reduce the amount of information necessary to represent a speech signal. The speech signal $s(n)$ is represented as a linear combination of p previous samples plus an error term

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e \tag{4.1}$$

Coefficients a_k are determined by minimizing the prediction error e.

In addition to providing an efficient speech coding framework, linear prediction of speech also corresponds to a simple speech production model, i.e. the convolution of an excitation by an all pole filter whose transfer function is of the form

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{4.2}$$

where $S(z)$, $H(z)$ et $U(z)$ are the z transforms of respectively the signal, the vocal tract filter and the excitation.

There exist efficient methods (see [232]) to compute the a_k coefficients. Using Eq.4.2, Eq. 4.1 can be reformulated to relate to the excitation signal $u(n)$ as

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n)$$

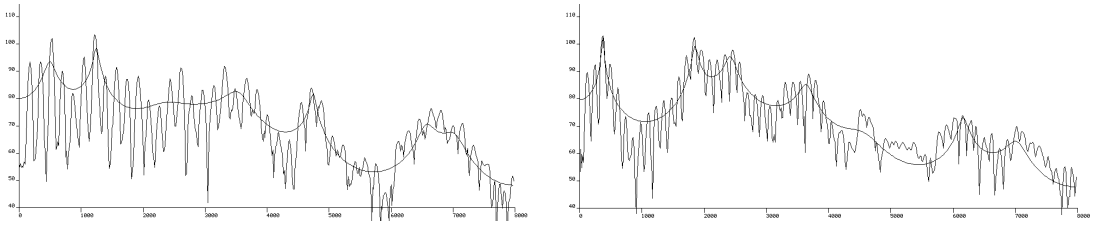


Figure 4.1: Narrow band Fourier transform and LPC spectra of /ɛ̃/ (left) and /ɛ/ (right)

Fig. 4.1 illustrates the weak and strong points of the linear prediction analysis. First, the optimization process, i.e. minimizing the error term, tends to adjust spectral peaks at harmonics which, of course, do not necessarily correspond to formant frequencies. Second, the underlying all pole filter approach does not enable spectra of nasal vowels or more generally nasalized sounds to be approximated correctly. As it clearly appears in Fig. 4.1 there is a strong spectral mismatch between 2,000 and 4,000 Hz for the nasal vowel /ɛ̃/ whereas there is a good fitting between LPC peaks and spectral peaks of /ɛ/. This default can be partly compensated by increasing the prediction order with the risk of recovering many spurious peaks.

The strong point of the linear prediction analysis is that it has a good spectral resolution. This is particularly important when two formants are close together (F1 and F2 of /u/, F2 and F3 of /y/, F3 and F4 of /i/).

Selective linear prediction [4] is interesting because it enables the prediction to focus on a part of the spectral domain. Solving Eq. 4.1 requires the calculation of autocorrelation coefficients, i.e. $\Phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k)$ where $s_n(m) = s(n+m)$. The underlying idea of selective LPC is to compute autocorrelation coefficients from the magnitude spectrum (see [233] pages 556 and following for instance). The main advantage is to get a better fitting over the spectral domain to analyze (see Fig. 4.2).

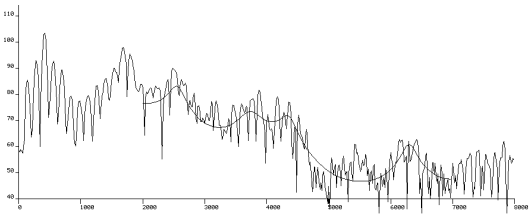


Figure 4.2: Selective linear prediction applied to the interval [2000Hz, 7000Hz].

Another approach derived from linear prediction is the **Line Spectrum Pair** representation (LSP) introduced by Itakura [234]. Line spectrum representation of linear prediction coefficients is mainly used in speech coding and synthesis because their filter stability preservation property enables quantization and interpolation.

The idea is to decompose the linear prediction polynomial $A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k}$ into two

polynomials $P(z)$ and $Q(z)$

$$\begin{array}{ll} P(z) & = A(z) + z^{-(p+1)}A(z^{-1}) \\ & = 1 + \sum_{k=1}^{p}(a_k + a_{p+1-k})z^{-p} + z^{-(p+1)} \\ Q(z) & = A(z) - z^{-(p+1)}A(z^{-1}) \\ & = 1 + \sum_{k=1}^{p}(a_k - a_{p+1-k})z^{-p} - z^{-(p+1)} \end{array} \quad (4.3)$$

$P(z)$ is a symmetric polynomial and $Q(z)$ is an anti-symmetric polynomial such that $A(z) = \frac{P(z)+Q(z)}{2}$. Finding roots of these two polynomials relies on the use of Chebyshev polynomials [235]. First, $P(z)$ and $Q(z)$ are transformed into Chebyshev polynomials. Then, roots of these polynomials are calculated through numerical algorithms.

This representation has three main properties:

- All the roots of $P(z)$ and $Q(z)$ are on the unit circle.
- The roots of $P(z)$ and$Q(z)$ are interlaced.
- $P(z)$ corresponds to the vocal tract with the glottis closed and $Q(z)$ to one with the glottis open.

LSPs have been used in a number of studies on the correlation between speech acoustics, vocal tract configuration and facial data [17–19] and multimodal speech synthesis [236].

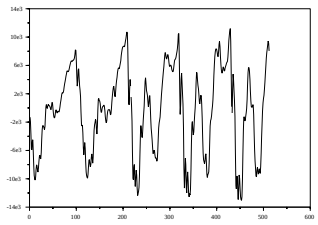
4.1.2 Cepstral smoothing

As acoustic-to-articulatory inversion focuses on the contribution of the vocal tract it is important to get a spectral analysis that removes the effect of the speech source in the spectral representation. From this point of view cepstral smoothing is a very good candidate since its principle is to separate source $e(n)$ and vocal tract contributions $h(n)$ of the speech signal $s(n)$ supposed to be the convolution of both: $s(n) = e(n) * h(n)$

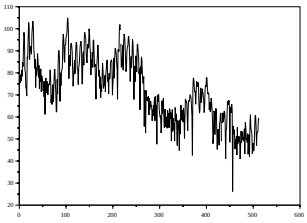
The principle of the cepstral analysis is a homomorphic processing that transforms the convolution into a simpler operation, i.e. a sum, that also clearly separates both contributions. The application of an inverse Fourier transform to the log magnitude spectrum of speech separates these two contributions quite well (see [232] for further details). The resulting vector is called cepstral coefficients (see Fig. 4.3(c)). Low order cepstral coefficients represent the “slow” variations of the spectrum shape, i.e. vocal tract, and high order coefficients “fast” variations, i.e. harmonics. A simple filtering, called liftering because it operates on cepstral coefficients, consisting of keeping only the first coefficients (see Fig. 4.3(d)) allows the contribution of the vocal tract to be isolated.

A further application of the Fourier transform enables a smooth spectrum to be obtained (see Fig. 4.3(f)).

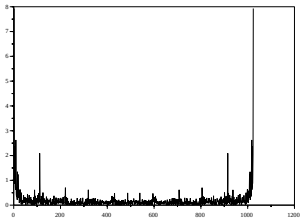
Unlike linear prediction coding the cepstral analysis does not impose any assumption about the analytical form of the vocal tract filter. This constitutes its main advantage since all speech sounds can be analyzed with the same parameters without adaptation.



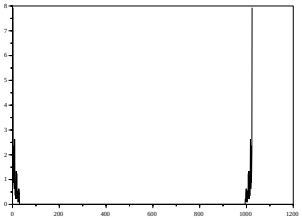
(a) speech signal



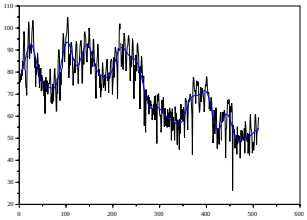
(b) log magnitude spectrum



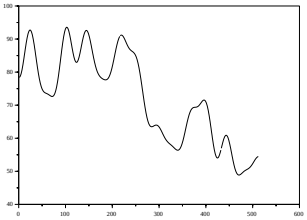
(c) cepstral coefficients (all but the first)



(d) cepstral coefficients liftered



(e) FT and smoothed spectra



(f) smoothed spectrum only

Figure 4.3: Defining cepstral coefficients for a speech signal, using cepstral smoothing, filtering and Fourier Transforms

4.1.2.1 Improving the approximation of spectral peaks

Cepstral smoothing approximates the whole spectrum whereas peaks of harmonics should be given a higher importance since their contribution to perception is larger than that of other spectral parts. Imai and Abe [237] thus proposed a very interesting improvement that consists of iterating the cepstral calculation on the part of the narrow band spectrum which is above the cepstrally smoothed spectrum.

Let S be the narrow band spectrum,
$V^{(1)} = \hat{S}$ ($\hat{S}$ is the cepstrally smoothed spectrum),
$E^{(1)} = g(S - \hat{S})$ where $g(y)$ = if $y > 0$ then y else 0,
$E^{(1)}$ represents the positive difference of S above $\hat{S}$,
$\hat{E}^{(1)}$ is the cepstral smoothing of this difference (see Fig. 4.4) which is added to $\hat{S}$ so to move the smoothed spectrum towards peaks.

The algorithm is based on iterations using an initial solution $\hat{E}^{(1)} = \sum_{m=0}^{N-1} e_m^{(1)} h_m cos(\frac{2}{N} mk)$, where N is the number of points of the Fourier transform, $e^{(1)} = IDFT(E^{(1)})$ and h_m is the liftering window.

For each iteration $i + 1$:
$V^{(i+1)} = V^{(i)} + \hat{E}^{(i)}$
$E^{(i+1)} = g(E^{(i)} - (1 + \alpha)\hat{E}^{(i)})$ where α is an acceleration factor.
$\hat{E}^{(i+1)} = DFT(h(IDFT(E^{(i+1)})))$
where $V^{(i)}$ is the envelope obtained at the previous iteration , $E^{(i)}$ and $\hat{E}^{(i)}$ are the positive difference and its smoothing.

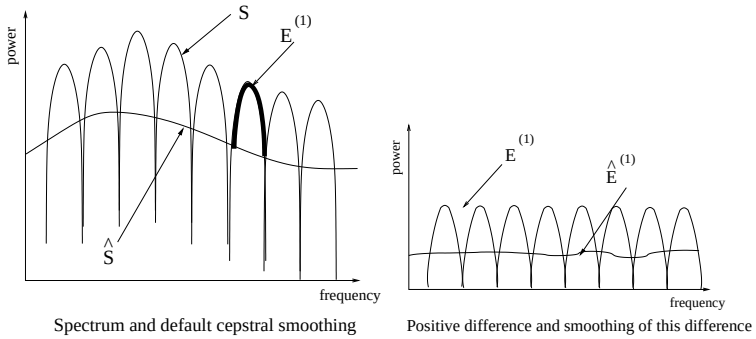


Figure 4.4: Principle of the true envelope calculation

Fig. 4.5 illustrates qualities of the true envelope compared to the simple cepstral smoothing. First, as expected, the true envelope fits harmonics better, but it has other qualities as well:

- The energy of peaks is in good agreement with that of harmonics contrary to peaks of the simple cepstral smoothing (one can note that the first peak of the cepstral smoothing is below the second, what does not reflect the energy of the corresponding harmonics).
- The stability and relevancy of true envelope peaks, which mainly correspond to formants, is better with true envelope than with default cepstral smoothing.
- The discrete cepstra method proposed by Gallas and Rodet [238] also approximates spectral peaks. However, the discrete cepstra method requires the knowledge of spectral peaks which thus have to be detected beforehand. The true envelope method prevents errors due to the detection of spectral peaks. The disadvantage is a larger computational cost since several iterations (six iterations are generally sufficient) are necessary, with each iteration corresponding to two Fourier transforms.

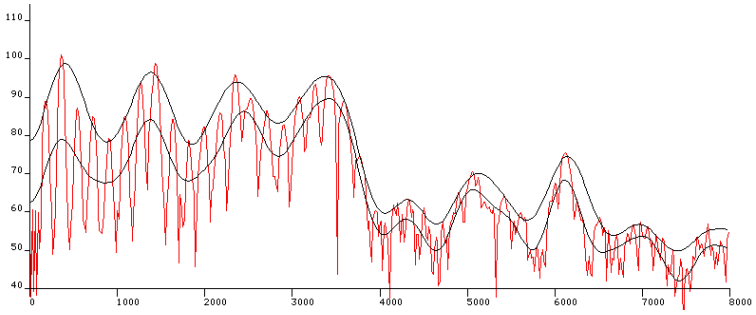


Figure 4.5: Narrow band spectrum, cepstrally smoothed spectrum (below harmonics) and true envelope (upper smoothed curve).

4.1.2.2 Mel cepstral analysis and perceptual frequency scales

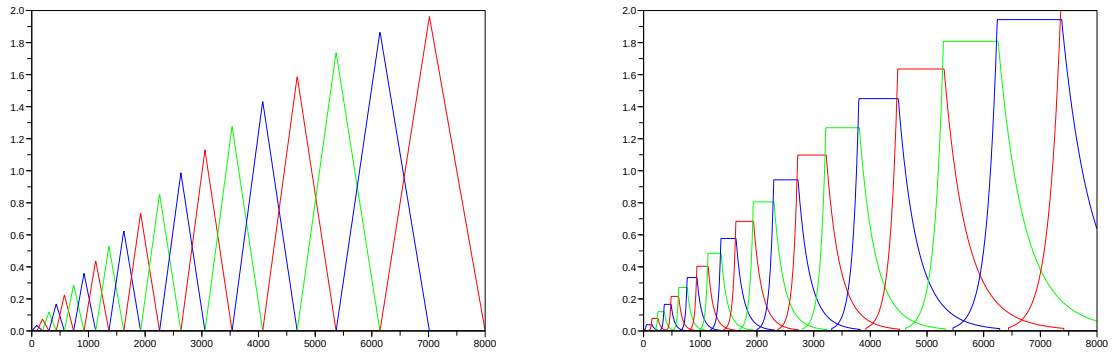
As the perceptive contribution of energy depends on the frequency, Davis and Mermelstein [239] proposed to apply a filtering that approaches the frequency resolution of the human auditory system. They thus use the Mel scale

$$f_{Mel} = \frac{1000}{log2} log(1 + \frac{f_{Hz}}{1000})$$

which gives approximately a linear frequency scale up to 1000 Hz and then a logarithmic scale, and they designed a filterbank in this scale (Fig. 4.6(a)). These filters are applied to a narrow band spectrum, then a discrete cosine transform (DCT) is applied to their outputs, denoted X_k:

$$MFCC_i = \sum_{k=1}^{N} X_k cos[i(k - \frac{1}{2})\frac{\pi}{N}]$$

The Mel cepstral coefficients $MFCC_i$ are very popular in automatic speech recognition because they give the best recognition rates for a wide range of applications. Generally, the window is between 20 and 32 ms long, 24 Mel filters are applied and the first 12 Mel cepstral coefficients are kept.



(a) Mel filters

(b) PLP critical bands

Figure 4.6: Mel filters and PLP critical bands (including the preemphasis).

As several attempts of acoustic-to-articulatory inversion use the Mel cepstral analysis it is important to get a better understanding of its properties with respect to the formant information which is crucial for inversion.

The first point we will address is the effect of the frequency scale and corresponding filters used to implement the perceptive scale. None of the perceptive filters (Mel and Bark) remove harmonics in low frequencies, as shown in Fig. 4.7. The smoothing effect of Bark filters is stronger than that of Mel filters because Mel filters are sharper at their maxima (a triangle instead of a plateau), (c.f. Fig. 4.1.2.2). However, the harmonics are kept by the Bark scale up to a higher frequency than by the Mel scale.

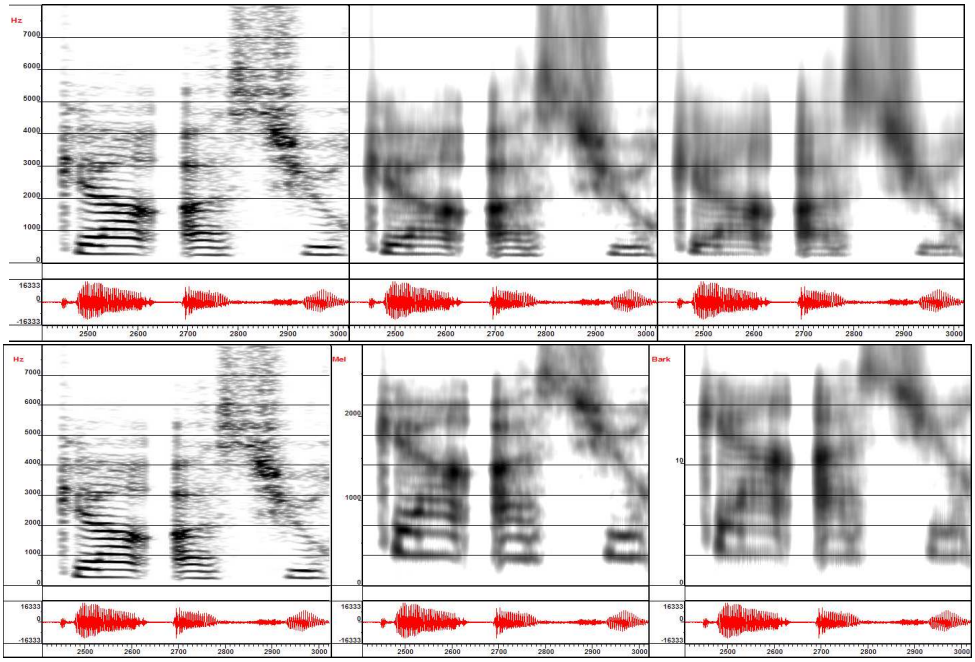


Figure 4.7: Narrow band (left), Mel filtered (center) and Bark filtered (right) spectrogram for a female speaker. In the upper row, the frequency unit is Herz, in the lower it is that of the analysis (respectively Herz, Mel and Bark).

The second point concerns the frequency resolution of the MFCC analysis. In order to evaluate the nature of the spectral information provided by the MFCC analysis an inverse discrete Fourier transform has been applied to Mel cepstra to obtain the corresponding spectrum. This spectrum is not the information used by automatic speech recognition for instance but it enables a clear evaluation of the formant structure and spectral information conveyed by MFCC.

The first column of Fig. 4.8 presents the effect of the number of filters. It turns out that formants are kept even if the number of filters is as low as 20 for the vowel /ε/ even if this would not be true for a vowel for which the second and third formants are closer.

The middle column presents the effect of the number of coefficients kept. In order to limit interferences with the number of filters a fairly high number of filters, i.e. 128, has been chosen since it corresponds to a very small degradation of the spectral information. With 32 or even 16 coefficients the first three formants are still visible. However, with 12 coefficients, there are three peaks left below 4kHz but with a valley instead of the third formant peak. The spectral information is thus not relevant anymore, at least within the context of acoustic-to-articulatory inversion. The third column presents the effect of the number of filters by keeping the number of coefficients set to 16. One can see that, even with a fairly small number of filters, good spectral information is kept. In short, the number of coefficients kept is the determining factor of the spectral quality of the Mel cepstral analysis. Unlike automatic speech recognition where the objective is to limit the amount of data necessary to build acoustical models of sounds, acoustic-to-articulatory inversion thus requires a higher number of coefficients to be kept.

Fig. 4.9 show the influence of the fundamental frequency on the Mel cepstral analysis. Results are presented, on the one hand for 32 filters and 16 coefficients kept, and for 24 filters and 12 coefficients kept on the other. The latter choice is very often used in automatic speech recognition. Contrary to the case of the male speaker the harmonics of the female speaker are not removed by Mel filters. Traces of harmonics indeed remain in spectrum of the female speaker, particularly the second one, whatever the number (16 or 12) of coefficients (right column). Globally, the spectral information is fairly less relevant than for the male speaker spectrum where the first three formants can be seen very clearly. The definition of a strategy for spectral analysis thus seems much more difficult because of the strong influence of the higher fundamental frequency.

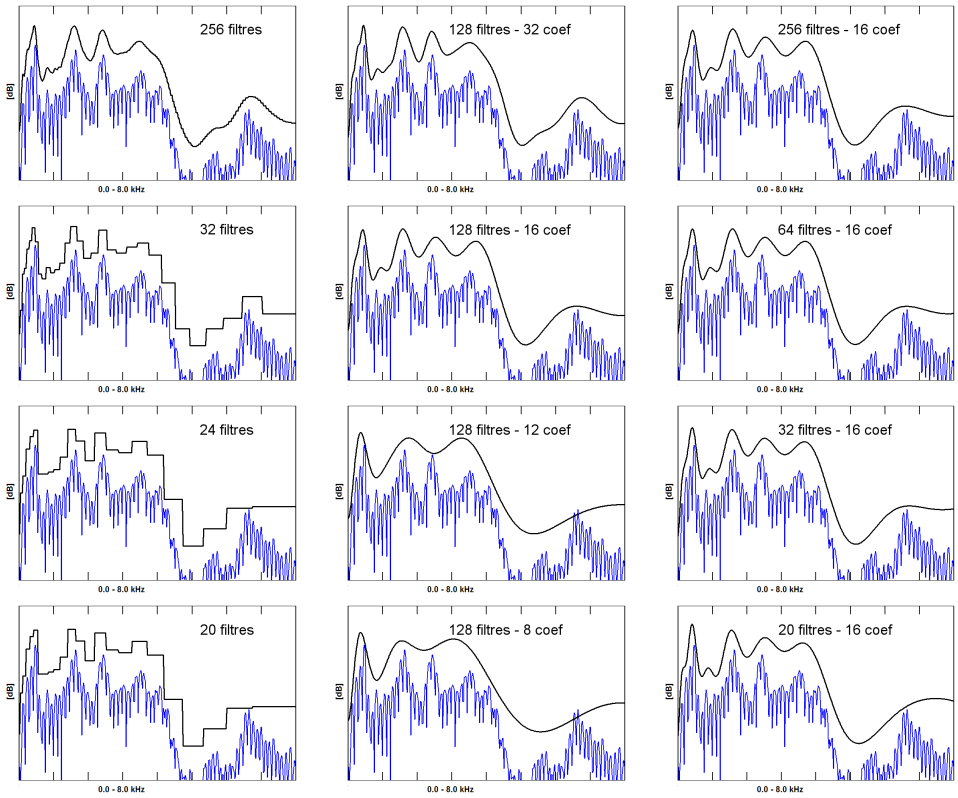


Figure 4.8: Effect of the number of Mel filters (left column), of the number of coefficients kept by the liftering (middle column) and of the number of filters while keeping the number of coefficients constant (right column). The curve below the MFCC smoothing is a narrow band Fourier transform over the same 32 ms window.

4.1.2.3 Perceptual linear prediction

Perceptual linear prediction was introduced by Hermansky [240] and consists of applying a linear analysis to the output of a critical band filter bank. The expected advantage is that copying the human auditory process gives rise to a spectral representation that is more robust to speaker variability and captures relevant acoustic features. Mimicking the human auditory

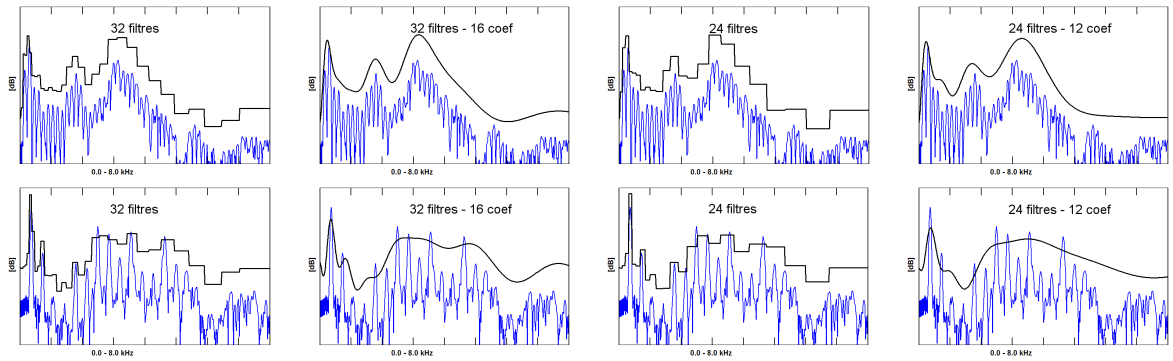


Figure 4.9: Application of Mel filters and Mel cepstral smoothing for a male (upper row) and a female (lower row) speaker (vowel /i/).

process corresponds to the use of the Bark scale ($Bark(f) = 6log((f/600) + \sqrt{(f/600)^2 + 1})$) instead of Mel scale for Mel cepstral analysis, the application of an equal loudness preemphasis and an intensity loudness power law ($.^{1/3}$) in order to reduce the spectral amplitude of the critical band output.

As for selective linear prediction, autocorrelation coefficients are obtained by applying an inverse Fourier transform to the perceptual magnitude spectrum. This enables the calculation of the linear prediction coefficients.

It should be noted that the preemphasis function used by Hermansky ($E(\omega) = (\omega^2 + 56.81010^6) * \omega^4/((\omega^2 + 6.31010^6)^2\omega * (\omega^2 + 0.381010^9))$ is actually very close to the traditional preemphasis (obtained by the differentiation of the speech signal) use to boost the speech signal spectrum. Therefore, PLP spectra shown in the following figures have been obtained with the traditional preemphasis.

Figs. 4.10 and 4.11 present the comparison between PLP and Mel cepstral smoothing for the same number of coefficients. The prior spectral filtering analysis is the Bark critical filter bank for PLP and the Mel filter bank for Mel cepstral analysis. Both analyzes capture almost the same spectral information. However, the effect of the intensity conversion tends to remove all spectral information above 2.5 kHz (making the spectral peaks disappear). It thus seems that intensity conversion should not be used in the framework of acoustic-to-articulatory inversion. Fig. 4.10 shows that the PLP (without intensity conversion) performs slightly better than the Mel cepstral analysis. Indeed, the F3 formant is better approximated, and more generally the spectrum below 4 kHz. Conversely, the Mel cepstral smoothing performs slightly better in high frequencies.

Fig. 4.11 shows spectra obtained with these two analyzes for a higher F0 speech. PLP favours low frequency spectral peaks, i.e. harmonics when F0 is high, more than Mel cepstral smoothing with the same number of coefficients. The higher the FO, the more PLP captures harmonics instead of formants, and therefore, the Mel cepstral smoothing seems more appropriate to perform acoustic-to-articulatory inversion.

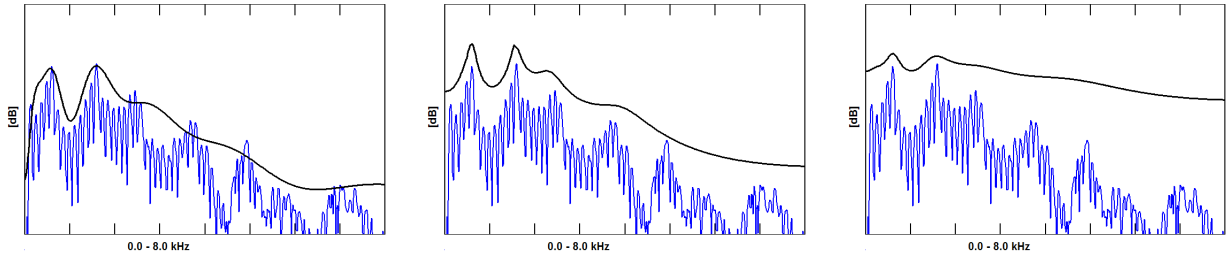


Figure 4.10: Comparison of Mel cepstral smoothing (24 bands, 12 coefficients), PLP (the prediction order is 12) without intensity conversion and with intensity conversion (from left to right) for a vowel /a/ pronounced by a male speaker.

4.1.3 Vocal Tract Resonance Tracking

Vocal Tract Resonances can represent speech efficiently and compactly in an intuitive way. Such a representation has been successfully exploited in various speech related areas such as synthesis [241], recognition [242] or speech inversion [243]. Accurate Vocal Tract Resonance tracking has been actively pursued by many researchers. Earlier efforts were based on spectral analysis and spectral peak-picking techniques [244–246]. Vocal Tract Resonances, or formants, largely coincide with prominences of the speech spectrum for non-nasalized vowels and semivowels and they have been generally regarded as such by traditional tracking algorithms. In cases, however, when the all-pole model for speech is not relevant, as with stops, fricatives and nasals, vocal tract resonances may not be directly observable as spectral peaks. This makes the tracking problem much more complex.

Deng et al. [247] propose a continuous-valued model for the resonances x (including the resonance bandwidths) that incorporates additional prior information in the form of hidden dynamics. Proper fusion with the observed speech acoustics is achieved in a Kalman filtering/smoothing framework. The target-directed state equation is:

$$x(k+1) = \Phi_{s(k)}x(k) + [I - \Phi_{s(k)}]u_{s(k)} + w_s(k) \tag{4.4}$$

where the state matrix $\Phi_{s(k)}$ and target vector $u_{s(k)}$ are considered to be phone-independent for simplicity. The target vector is a way to introduce prior nominal values for the resonances, e.g. $u = (500Hz, 1500Hz, 2500Hz, 80Hz, 120Hz, 150Hz)$ for three resonances, $P = 3$. The

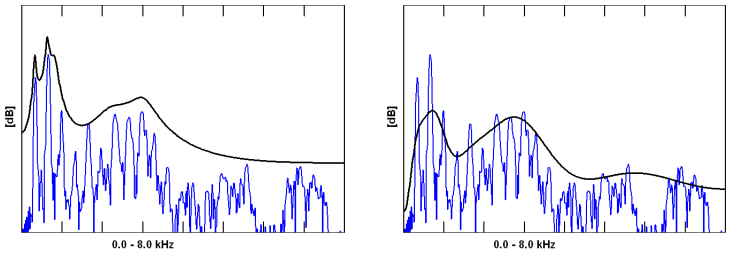


Figure 4.11: Comparison of Mel cepstral smoothing (24 bands, 12 coefficients) and PLP (the prediction order is 12) for a high F0 value.

observation vector o includes LPC cepstrum coefficients. Based on an all-pole speech model, these coefficients may be expressed as:

$$C(i) = \sum_{p=1}^{P} \frac{2}{i} e^{-\pi i \frac{b_p}{f_s}} cos(2\pi i \frac{f_p}{f_s}), \quad i = 1, \ldots, I \tag{4.5}$$

where f_s is the sampling frequency and f_p, b_p are the frequency and bandwidth of the p-th Vocal Tract Resonance. To account for errors due to zeros and additional poles beyond P, the prediction residual μ and the zero-mean noise $v(k)$ are introduced. The observation equation is:

$$o(k) = C[x(k)] + \mu + v(k) \tag{4.6}$$

which, though highly nonlinear, can be piecewise linearized as is demonstrated in [247]. Vocal Tract Resonance tracking is then achieved by an adaptive Kalman filter and smoother. The prediction residual is updated online so that the observation model best fits the current speech utterance. The overall algorithm is quite elegant and the results are promising. A similar observation model is used by Zheng and Hasegawa-Johnson in [248]. Phone-dependent information is also incorporated in a mixture-state particle filter framework.

Togneri and Deng in [249] present an extended Kalman filtering framework to track Vocal Tract Resonances from MFCC. The multivariate and nonlinear observation equation is implemented by multiple switching MLP (multi-layer perceptron) neural networks. The parameters of the model (including the MLP weights) are trained using the Expectation-Maximization algorithm and formant estimates as given by a conventional formant tracker. The training process is what mainly differentiates this approach from previous work by Deng and Ma [242]. In [250], Deng et al. apply a discrete-value approach by quantizing the Vocal Tract Resonance space and then use the Viterbi algorithm to find the optimal tracks.

Toledano et al. [251] extract candidate formants based on Linear Prediction analysis and then find the best trajectories using properly initialized and trained phone HMMs. Context-dependent phone HMMs and training initialization based on manual formant trajectory labeling are found to give the most satisfactory results. Context-dependent phonemic information is also exploited by Lee et al. in [252]. The idea is to first extract nominal formant trajectories based on the phoneme sequence in the given speech utterance and then interpolating applying certain coarticulation rules. Then candidate formants are estimated by linear prediction and dynamic programming is used to find the cost-minimizing tracks. The minimized cost function mainly penalizes divergence from the nominal values.

From a different viewpoint, Mustafa and Bruce [253] propose an algorithm mainly with noise and speaker robustness in mind. It is based on the decomposition of speech into modulated components [254]. The signal is first filtered by an adaptive bandpass filterbank including four formant filters. Each of them consists of one pole at the corresponding formant frequency and three zeros, one at each of the other three formant frequencies. If the energy at a certain band is above a threshold and the previous frame is voiced then a single pole model is fitted to the narrow band signal. The energy, voicing and gender detectors used are updated adaptively. In case the current formant estimates are not judged as reliable, then moving average values are used.

An interesting formant tracking algorithm is presented in [255] by Laprie. It extends earlier

work by Laprie and Berger [256]. The main idea is to find the formant tracks by minimizing a functional which guides the tracks to areas of high spectral density and further imposes smoothness. This functional is as follows:

$$E(F) = -\int_{t_i}^{t_f} E_{spec}(t, F(t))dt + \lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \tag{4.7}$$

The first term represents spectrogram energy along the formant track. The second term represents the track length and the curvature. For regular curves, this term should be small. Certain ways are suggested to incorporate interdependency between formants so that each formant won't evolve independently and so, among others, the initialization stage will be simple enough. The first strategy given is called the spectrogram partition strategy and assigns adaptively a spectrogram partition based on the current formant values. This partitioning is not probability-based and it does not work well in case of nasalized sounds, when a formant does not correspond to a spectral peak. Alternatively the repulsive track strategy may be applied according to which a new (exponential) term is added to the energy functional for minimization.

$$E(F) = -\int_{t_i}^{t_f} E_{spec}(t, F(t))dt + \mu \sum_n E_{spec}(t, F_n(t)) exp^{-(\frac{F_n(t)-F(t)}{s_n})^2} + \lambda \int_{t_i}^{t_f} \alpha|F'(t)|^2 + \beta|F''(t)|^2 dt \tag{4.8}$$

This term penalizes tracks that approach each other.

Alternative speech representations for formant tracking have been proposed in [257, 258]. In [257], Bozkurt et al. try to remove the effect of glottal source to the speech signal before estimating the formants. They try to make the peaks of the spectrum that correspond to vocal tract resonances more prominent. For this reason, they get the formant peaks on the differential phase spectrum. This is the negative derivative of the phase of the chirp-z transform spectrum of the signal. The chirp-z transform spectrum is actually estimated not on the unit circle but at radius r. No other specific speech model is used, only peak-picking.

Potamianos and Maragos in [258] have proposed the so-called pyknogram to extract initial formant candidates. In a Gabor multiband scheme, raw frequency and bandwidth measurements are extracted from each filter as first and second spectral moments of the signals at the output:

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)[a(t)]^2 dt}{\int_{t_0}^{t_0+T} [a(t)]^2 dt}, \quad B_w = \frac{\int_{t_0}^{t_0+T} [(\dot{a}(t)/2\pi)^2 + (f(t) - F_w)^2 [a(t)]^2] dt}{\int_{t_0}^{t_0+T} [a(t)]^2 dt} \tag{4.9}$$

where $a(t)$ is the instantaneous amplitude and $f(t)$ is the instantaneous frequency of the signal at the output of a filter. These are frame-based estimates and may provide an alternative speech representation, known as pyknogram and shown in Fig. 4.12.

By simple and robust thresholding only a limited number of these moments are kept, corresponding to areas of high spectral density. These are considered to be raw formant estimates and are shown in Fig. 4.12 superimposed on the pyknogram. The formant tracks that are derived using these raw formants and a dynamic programming algorithm are also shown in Fig. 4.12. The algorithm is looking for four formant tracks in each utterance. Initially, the algorithm proposed in [258] followed a rule-based approach to extract the final tracks, after [246]. Instead,

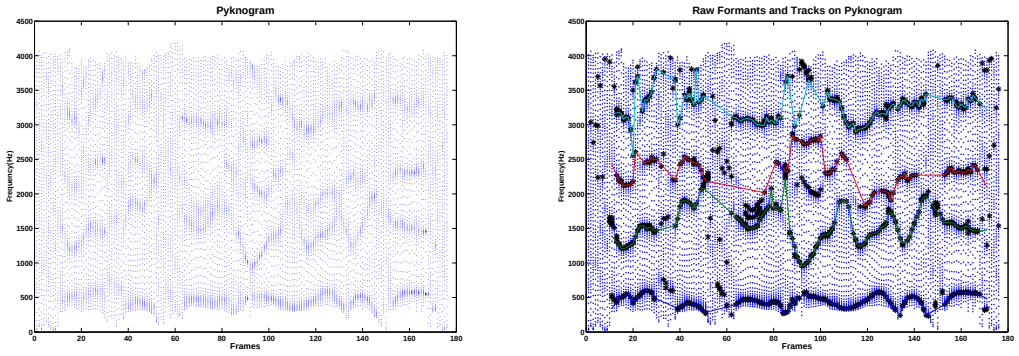


Figure 4.12: Pyknogram of the utterance "Don't ask me to carry an oily rag like that", with superimposed raw formants and formant tracks on the right.

to improve robustness, an alternative is proposed which is based on dynamic programming, similar to [259].

To find the best set of trajectories for the formants through a trellis of candidate mappings, see Fig. 4.13, the cost of mapping candidate frequencies to formants at each frame is minimized over all analysis frames. Each node corresponds to a different mapping. The number of nodes may be significantly reduced considering that each mapping has to satisfy certain rules, namely $F_1 < F_2 < \ldots$ and (with values in Hz):

$$100 \leq F_1 \leq 1500, \quad 500 \leq F_2 \leq 3500, \quad 1000 \leq F_3 \leq 4500, \quad 2000 \leq F_4 \leq 5000. \tag{4.10}$$

There could also be times when a formant estimate is unreliable and so it would be better to ignore it. To account for this, the null estimate $\emptyset$ is also considered as a possibility. So, for example, in case there is a set $\{450, 1300, 3400\}$ as raw formants for a certain frame, two of the possible nodes in this frame would correspond to the quadruples $\{450, \emptyset, 1300, 3400\}$ or $\{450, 1300, 3400, \emptyset\}$.

The Dynamic Programming cost function that should be minimized is defined on the trellis as:

$$C[t,n] = C_{local}[t,n] + \min_m \{C_{tran}[(t,n),(t-1,m)] + C[t-1,m]\} \tag{4.11}$$

where t is a frame index and n, m are node counters. Locally, formant estimates for which the corresponding bandwidth estimates are big should be penalized. Further, cases in which the formant estimates are closer to empirically predefined expected values E_f and have fewer null estimates should be favoured. So, the local cost at each node is defined as:

$$C_{local}[t,n] = \sum_i \alpha_i B_{w,i}^2 + \beta_i |F_i - E_f\{i\}|/E_f\{i\} + \gamma_i \delta_{F_i,\emptyset} \tag{4.12}$$

The $\delta_{F_i,\emptyset}$ in the last addend equals to unity if the formant estimate at the position i is null. It equals to zero elsewhere. The expected formant values are $E_f = \{500, 1500, 2500, 3500\}Hz$. For the transitions, big jumps between subsequent frames are penalized, and those involving null estimates should be less favoured. The transition cost is therefore:

$$C_{tran}[(t,n),(t-1,m)] = \sum_i \epsilon_i \Big(\frac{Fi(t,n) - Fi(t-1,m)}{F_{spike}}\Big)^2 + \zeta_i(\delta_{F_i(t,n),\emptyset} + \delta_{F_i(t-1,m),\emptyset}) \tag{4.13}$$

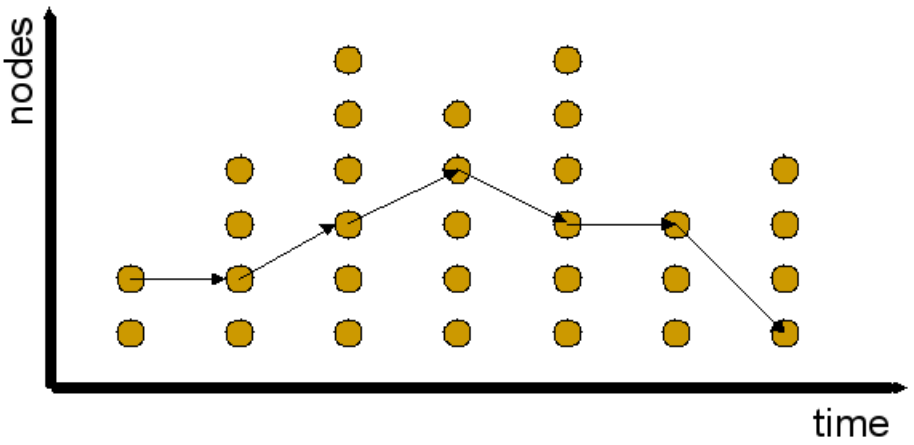


Figure 4.13: Trellis including all the possible mappings of raw formant estimates to the predefined number of formant positions for which we would like to have reliable values. A dynamic programming search in this trellis provides us the optimal formant tracks, based on the given raw formant estimates and the cost function we have defined.

$F_{spike} = 500Hz$ is the maximum jump allowed for the formant values in a track between subsequent frames. In Fig. 4.12, an example is shown of a set of formant trajectories derived using the dynamic programming approach along with raw formant estimates from the multiband energy demodulation algorithm by [258].

Chapter 5

Inversion Methods

5.1 Explicit inversion from acoustics to vocal shape

Mermelstein [2] studied the link between the eigenfrequencies and area function of a vocal tract that is completely loss-less. He showed that if the log-area is band-limited so as to preserve only *2n* Fourier cosine series coefficients, the *n* low-frequency poles and zeros of the admittance at the lips determine the cosine series coefficients exactly. The proof is formal for shapes that are small perturbations of the uniform tract. The extension to any tract shape is empirical. The series coefficients that cannot be determined via a measurement of the formant frequencies were assumed to be zero.

The formant frequencies are the observable spectral effects of the eigenfrequencies of the vocal tract, which correspond to the admittance poles, i.e. the frequencies for which a zero acoustic pressure coexists with a finite volume velocity at the lips. The admittance zeros, i.e. the frequencies for which a finite acoustic pressure coexists with a zero volume velocity cannot be extracted from the speech signal, because they describe the closed-lips condition during which no sound is emitted, e.g. [p]. Formant frequencies alone are therefore not enough to determine the area function unless a priori values are assigned to the even coefficients of the expansion of the cross-sections into a Fourier cosine series.

In an approach based on Mermelstein's proposal [12, 20], the log-area function was developed into a truncated cosine series and the series coefficients that could not be determined via formant frequencies were fixed so that the final shape was as close as possible to the uniform vocal tract. Another method in the same vein was the use of the distinctive regions and modes model, described in Section 3.1.2, for inverse mapping [10]. In another study, Yu [25] expanded the anterior part of the area function only, so as to be able to constrain the epilaryngeal end of the acoustic duct, to avoid anatomically unrealistic tract shapes.

A transfer-function to area transform that has played an equally seminal role as Mermelstein's proposal is based on linear predictive analysis of speech, described in Section 4.1.1. It has been shown that the regression coefficients can be mathematically turned into reflection coefficients that have an interpretation in terms of an area function model made up of a concatenation of cylindrical tubelets. The number of reflection coefficients is equal to the number

of regression coefficients. The cross-sections of the tubelets are determined algebraically from the reflection coefficients, up to a multiplicative constant.

The conditions under which the area function can be recovered exactly from the prediction coefficients were listed in the introduction, but in reality, human speech signals are not produced under conditions that enables recovery of the tract shape precisely and reliably by this method. It is, however, the case that linear prediction-based transfer-function to area mapping is one of the best known speech-to-shape transforms, which is explained in many text books. Recently, Krstulovic has proposed an extension that enables computing concatenated-cylinder area functions, the cylinders of which are of unequal lengths [260].

Hybrid methods that combine Mermelstein's with speech-to-shape inversion via linear regression of speech samples have been proposed by Wakita and Gray [261] and Mokhtari [262]. Wakita and Gray determine, in a first step, an area function via linear predictive analysis and then compute, in a second step, the open-lips admittance zeros and closed-lips admittance poles from the reclaimed shape. The virtual admittance zeros and poles are then used to uniquely determine the shape of the loss-less Mermelstein log-area model.

Mokhtari developed a hybrid approach that involves a critical study of the contribution of the formant bandwidths to the recovery of the area function via linear predictive coefficients. The recovered area function was represented by means of the odd-numbered cosine and sine terms. The sine coefficients are claimed to be closely related to the formant bandwidths and the cosine coefficients to the formant frequencies. The computed shapes have been used as a tool to study speaker-specific variability in speech sounds. The evaluation of the inversion method per se consisted in a visual inspection of the agreement between observed and reclaimed static vowel area functions.

The most often used general-purpose method for pseudo-inverting any matrix, which is not square or the determinant of which is zero or nearly zero is based on singular value decomposition [263]. Because, singular value decomposition is able to (pseudo)-invert linear relations only, its application to formant-to-area mapping involves locally linearizing the causal link between parameters and eigenfrequency before pseudo-inverting to obtain the morphological parameters as a function of eigenfrequencies.

Schoentgen and Ciocea [264] have shown that linear pseudo-inversion is a tool flexible enough to enable solving the formant-to-area mapping problem for static as well as evolving formant trajectories. The vocal tract model was a concatenation of cylindrical tubelets with time-varying cross-section areas and lengths. The time increments of the tubelet parameters have been obtained by inverting a linear algebraic system of equations that relate formant frequency and tract parameter increments. The elements of the matrix are estimates of the partial derivatives of the eigenfrequencies with reference to the model parameters. The increments are then added to previous cross-sections and lengths to recover their motion. Because more than one area function is compatible with the observed formant frequencies, pseudo-energy constraints have been used to determine a unique solution. The agreement between observed and model-generated formant frequencies has been better than 0.01 Hz. The method has been evaluated by computing the similarities between observed and computed static as well as evolving area functions [265].

Since the number of reasonably smooth area functions that agree with given acoustic data

are numerous, due to the weak constraints imposed on area functions, it may be more efficient to represent the vocal tract shape in a midsagittal articulatory model. It is, however, the case that not even anatomically correct models give rise to unique solutions when the acoustic data are the first few formant frequencies, which are the spectral cues that are phonetically meaningful [8]. It is also an open problem whether the parameters of a midsagittal articulatory model can be determined uniquely, once the area function is known exactly. The issue is not the heuristic that turns the two-dimensional sagittal profile into a three-dimensional area function, because the heuristic is selected so as to be invertible. The problem rather is whether articulators can be positioned so that different postures give rise to (nearly) identical midsagittal cross-sections. The answer presumably depends on the sophistication of the articulatory model and has, as far as we know, not yet been investigated thoroughly. In practice, authors use either articulatory or anatomically-constrained midsagittal profiles, or anatomically-constrained area functions. The purpose of the use of models is to decrease the number of parameters that are free to vary and constrain the parameters and their evolution to be anatomically and physiologically plausible.

5.2 Inversion-by-synthesis

As opposed to the ”explicit” inversion described in the previous section, ”implicit” inversion designates the iteration of synthesis model parameters until the observed and modeled acoustic data agree. The causal link between shape and acoustic data is thus not inverted explicitly. Instead, the model is assimilated to a plant the morphological input parameters of which are manipulated so as to optimize the acoustic output. The output is considered optimal when it agrees with the observed acoustic data. The acoustic data may be formant frequencies or whole spectra. Additional constraints that are routinely used pertain to the spatial and temporal smoothness of the area function, as well as its distance from the neutral tract or its kinetic pseudo-energy.

Two types of synthesizers may be used, based on an area function model or on an articulatory model, with the latter being more common.

An **o**ptimization-based transfer function to area conversion was developed by Flanagan, Ishizaka and Shipley [14]. The area function model comprised six parameters and was able to mimic the tract shapes of Russian vowels, obtained by X-rays [200]. The synthesizer was based on a temporal simulation of the lossy wave propagation within the vocal tract, with a two-mass model to simulate voicing, latent noise sources distributed along the vocal tract to generate turbulence noise for fricatives and plosives and a nasal tract. The cost function was the squared difference between the log-amplitude spectra over an analysis interval. Smooth evolution of the parameters as well as rate-of-change constraints have also been imposed. A difference between cepstral maxima is used to control the combined tension/mass parameter of the two-mass model of the vocal folds.

The starting values for a multi-parameter optimization were fixed by measuring the mouth area optically and changing the model parameters independently so as to minimize the cost function, using neutral values for other parameters. The evaluation of the method was performed by means of artificial shapes as well as an [ai] transition spoken by one of the authors.

An **o**ptimization-based transfer-function to articulatory model inversion, inspired by Flana-

gan et al. was reported by Levinson and Schmidt [41]. It consists a midsagittal articulatory model [266] instead of an area function model. The vocal tract model is lossy and comprises spectral models of glottal source and radiation load. The cost function includes the distance between the log-magnitude of the model transfer function and the estimated spectral envelope of the speech signal. The envelopes are extracted pitch-synchronously. The cost function is minimized by means of a steepest-descent algorithm. The articulatory parameter array for which a minimum is reached is the desired articulatory configuration.

When evaluating static synthetic vowels, the spectral agreement was typically 2 dB. When tracking diphthongs, the spectral match was 3 dB, but some articulators occasionally tended to "freeze" and the spectral matching was then performed by the other articulators in "ventriloquist" fashion.

In a similar spectral-to-articulatory inversion scheme [39, 267], the articulatory model involved morphological mimics of the jaw, velum, pharynx, as well as a statistical tongue model, inferred from X-ray data. The inversion was performed by means of a "hill-climbing" optimizer that minimizes the distance between observed and synthetic cepstra. In addition, the distances between the reclaimed and neutral tract shapes, as well as between the previous and present shapes have been minimized. The weights of the constraints were fixed manually.

The evaluation of the inversion method was based on a visual inspection of the agreement between computed and observed static Japanese vowels as well as the smoothness of the evolving parameter trajectories. One quantitative test has been the automatic recognition of five Japanese vowels spoken by 15 speakers.

Later Shirai and Kobayashi [268] have compared inversion by means of optimization and artificial neural nets [37]. They concluded that inversion based on a single feed-forward artificial neural net is not suited to the task of spectral-to-articulatory inversion. This observation has been confirmed by Rahim [6].

A formant-to-articulatory inversion method has been investigated by Sorokin [15]. The articulatory model describes the vocal tract by means of the tabulated surfaces of the pharynx, velum, hard palate, lips and tongue measured on X-ray film, as well as a 15-parameter model of the midsagittal and frontal profiles.

The cost function involves a minimization of the "muscle work":

$$W(z) = \Sigma_{i=1}^{N} c_i (z_i - z_i^0)^2,$$

in which N is the number of articulatory parameters, z_i is the i^{th} articulatory parameter, z_i^0 its neutral position, and c_i is the elastic resistance to a change of the articulatory parameter. Sorokin obtained the values of c_i through physiological experiments on the corresponding articulatory organs. The value $W(z)$ may be considered as the potential energy associated with the problem. Formants were compared on a logarithmic scale, a choice which appears to be perceptually inspired, and the distance between modeled and target formants was multiplied by a weight that changes from low to high during optimization.

Sorokin argued that an optimizer must be used that only involves the coordinates of the articulatory parameters, and not their rates of change. The optimizer stops once the distance is smaller than a critical threshold, which is frequency-dependent. Also, for vowels, any cross-section smaller than 0.3 cm^2 was rejected, because audible turbulence noise is generated for

smaller apertures.

The evaluation has been based on six Russian vowels [200] as well as on microbeam data of a male and a female speaker producing [bVbVbV] sequences. When up to four formants were used, formant frequency errors were typically a few percent, but could be as high as 20 percent. The explanation of the imperfect fit between observed and modeled formants seemed to be the optimizer, the manual initialization of which by means of the neutral tract shape appeared to be inadequate. It was also shown that four reference formants did not have an advantage over three. Sometimes the fourth formant made the fit worse because of measurement errors. Similarly, the inclusion of the formant amplitudes led occasionally to a poorer match between reclaimed and observed shapes.

Later, the same author has attempted to recover the tract shapes of a speaker producing unvoiced Russian fricatives [269]. The optimization method and cost function have been similar to those used previously. The formant frequencies have been replaced by whole-spectra, however. The match between observed and computed spectra was expressed via the inter-correlation coefficient [15]. The main conclusion has been that convergence to actual shapes is only observed when the optimization has been initialized manually.

The judgment of the agreement between observed and recovered vowel tract shapes was visual [15]. The shapes shown in the article demonstrate that the tongue profiles have not been recovered exactly; other articulators have not been evaluated. Methods reported in [15] and [269] have been evaluated on corpora larger than usual and they discuss the recovery of original shapes.

Generally speaking, difficulties experienced by authors who have investigated acoustic-to-shape inversion by optimization, have motivated the development of codebook-based methods, described in Section 5.3. These enable homing in on all approximately acceptable tract shapes or articulatory configurations for a set of observed acoustic data. The approximate shapes can then be used to initialize an optimizer that refines the codebook entries until the modeled and observed acoustic data agree [35] [34] [270]. Indeed, the main problem that confronts inversion by optimization appears to be local minima. The heterogeneity of the cost functions is a possible explanation.

5.3 Codebook methods

For a stationary vocal tract, the articulatory-acoustic mapping can be represented as a multidimensional function of a multidimensional argument: $y = f(x)$, where x, y are vectors describing the vocal-tract shape and the resulting acoustic output, respectively. In this section, we will review methods that exploit tables (also called *codebooks*) of precomputed couples (x,y) organized in a way to easily recover several articulatory vectors from a given acoustic vector.

Acoustic-to-articulatory inversion using codebooks has been studied for a long time, and the work of Atal et al. [5] constitutes to this regard a fundamental and remarkable work.

5.3.1 Fundamentals

The work by Atal et al. [5] was the first to investigate the use of codebooks for inversion. Although part of it is a little outdated (in particular the amount of data in the codebook) most of the theoretical work is still valid, and is actually still used in recent works. In this subsection, we will summarize the main results of this work and show how more recent studies derive from it.

5.3.1.1 Function inversion using codebooks

Inverting a multidimensional function using codebooks is conceptually simple. It consists of calculating $y = f(x)$ for a large number of different values of x, and of organizing the resulting pairs y, x based on the vector y. Finding a value of x corresponding to a given y consists simply of looking up the desired y in the data and obtaining the x associated with it. Some further complications arise when one has to deal with ambiguities and interpolation.

Point ambiguities are simply handled by organizing the codebook depending on the values of the y vectors: when two or more points in the x space produce identical, or nearly identical, values of y, these values will be placed in the same "region", or in neighboring regions.

Depending on the nature of the articulatory-to-acoustic function f, and particularly on the number of dimensions in the articulatory and acoustic spaces, ambiguities of a continuous nature may exist, that is, an entire subspace which maps onto a given acoustic point y. Those regions in the x space which produce no change in y, are called *fibers* by Atal. A fiber thus determines vocal tract shapes having identical acoustic properties. It is, in general, difficult to treat ambiguities of nonlinear functions, and the articulatory-acoustic relationship is not necessarily linear. However, if f is sufficiently well behaved to be linearized locally, the ambiguities of the linearized function can be characterized and studied. Using computational methods to extend the linarized regions in small steps, it is thus possible to systematically explore the entire non-linear regions.

We denote m the dimension of the x space, and n the dimension of the y space. In the vicinity of a point x_0, $y_0 = f(x_0)$ is approximated by y

$$y \approx y_0 + B(x - x_0),$$

where B could be the Jacobian matrix of f, that is, the matrix of partial derivatives of f. Specifically, if b_{ij} is the element at the i^{th} row and the j^{th} column of B, then

$$b_{ij} = \frac{\partial f_i}{\partial x_j}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, m,$$

could be the partial derivative of the i^{th} component of the vector $f(x)$ with respect to the j^{th} component of x. B is thus a $n \times m$ matrix. In practice, B is evaluated by approximating partial derivatives with partial differences.

5.3.1.2 Acoustic simulation

We describe here the method used by Atal et al. to compute acoustic parameters for a given vocal tract shape. Formants were used to represent the speech signal because this leads to an easy physical interpretability of the results.

The acoustic simulation itself is fairly classic: The vocal tract is regarded as a nonuniform acoustic tube terminated by the glottis at one end and by lips at the other end. It is assumed that:

- the vocal tract is excited at the glottis and the sound is radiated at lips,
- the vocal tract has no side branches or cavities,
- the cross-sectional dimensions of the vocal tract are small compared to a wavelength in the frequency range of interest,
- only plane waves propagate in the vocal tract.

As usual, the numerical computation is simplified by approximating the continuously varying area function by a series connection of a large number of uniform tubes. For plane-wave propagation, pressure and volume velocity are continuous functions of the axial dimension in the tube; the transmission matrix of the entire tube can thus be expressed as a product of such matrices for each uniform section. Atal et al. also take into account five principal sources of energy loss in the vocal tract: (i) viscous loss in a boundary layer at the surface of the tube, (ii) heat conduction loss at the vocal tract walls, (iii) radiation loss at the mouth opening, (iv) energy loss due to yielding of the vocal tract walls, and (v) energy loss at the glottis. The losses in the tube are represented by lumped elements at the input of each uniform section.

Let $F(s)$ be the velocity transfer function of the vocal tract at complex frequency s ($s = \sigma + j\omega$), defined by the ratio of the volume velocity at the lips to the volume velocity supplied by the glottal source, both defined as a function of the complex frequency variable s.

$F(s)$ can be calculated for a given area function specified in terms of cross-sectional areas and lengths of each uniform section. Then the formant frequency bandwidths are determined by finding the poles of $F(s)$ (i.e. the zeros of $1/F(s)$).

5.3.1.3 Vocal tract models

There are many different ways of describing the vocal tract shape, the most straightforward description being in terms of a number of cross-sectional areas specified at equidistant points from glottis to lips. Such a description however is quite inefficient: if the area function of the vocal tract is sampled every 0.5 cm, it means that 34 areas must be specified for a vocal tract 17 cm in length. A more meaningful way of describing the vocal tract is to specify the area function in terms of a few articulatory variables representing positions of different articulators, as described in Chapter 3. Atal et al. used an extension of the articulatory model of Stevens and House [271] in which the area function is described by four articulatory variables.

5.3.1.4 Experiments

The articulatory space was sampled along a four-dimensional grid. Compared to today's standard, the resulting codebook was fairly small: the sampling used resulted in a total of 30720 different vocal tract configurations, for which the frequencies, bandwidths, and amplitudes of the first five formants were computed. The y vector however only included the frequencies of the first three of these formants.

To organize the data, the three-dimensional acoustic space was partitioned into "cubes" (parallelepipeds, actually) of equal volume. The coordinate axes corresponding to the first two formants were divided into 50 Hz intervals while the axis for the third formant was divided into 100 Hz intervals. The data was then sorted into the cubes according to the frequencies of the three formants. A "cube" in general has many entries which are scattered due to the finite quantization of the y space. Atal et al used an iterative procedure to move y from each point to the center of the cube, exploiting the local linearity of the articulatory-to-acoustic function; all of the corresponding points in the cube are thus modified to merge into a single point at the center.

Atal et al. exploited these data in different ways; one of the most interesting result was the study of the one-dimensional fibers for eight non-nasal vowels. It was the first illustration of the non-uniqueness of the acoustic-to-articulatory mapping.

5.3.2 Codebook inversion of speech sequences

Among the many different approaches to the acoustic-to-articulatory inversion problem, most of them, implicitly or explicitly, use codebooks. The work of Atal et al. [5] constitutes a starting point, but unfortunately it did not study dynamic inversion, that is, the inversion of speech sequences, but only isolated vowels.

Most recent approaches of "traditional" (that is, using an explicit codebook) dynamic inversion proceed in 3 steps:

1. For each acoustic vector, a number of articulatory vectors are generated using a codebook look-up procedure (see Fig. 5.1).
2. An *initial* articulatory trajectory is derived from these vectors using e.g. dynamic programming and regularity constraints (see Fig. 5.2). This initial trajectory is a starting point for further optimization: the selection of a trajectory from a whole discretized articulatory space makes it easier to avoid local minima.
3. This trajectory is refined using e.g. optimization or variational regularization to obtain a smooth trajectory with a better acoustic accuracy. The solution obtained can be much smoother and acoustically faithful since the articulatory vectors are now able to vary in the continuous articulatory space.

In some cases however, neural networks methods for instance, dynamic inversion can be performed in a single step, depending on the data trained.

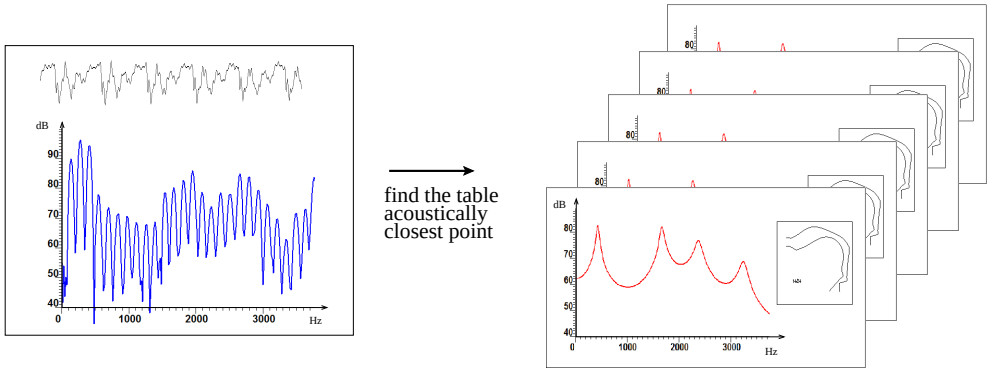


Figure 5.1: Searching for articulatory points that can correspond to a speech spectrum. Table points are represented at right and best fitting points are found out from the spectral information or from the frequencies of the first three formant frequencies.

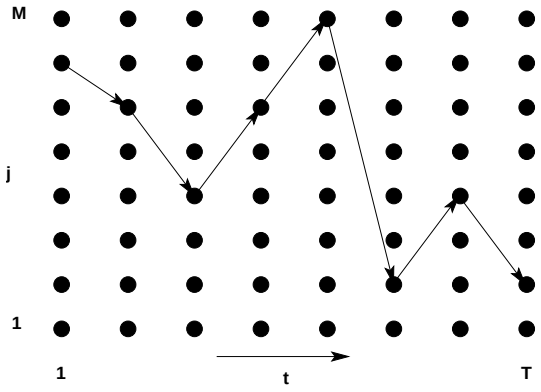


Figure 5.2: Searching for the best articulatory trajectory through dynamic programming. Each column gives articulatory points recovered from acoustic data at time t. M is the number of points in the articulatory table. T is the duration of the speech signal to invert. Arrows give a trajectory example.

A crucial preliminary stage is the construction of the acoustic-to-articulatory codebook which consists of selecting a method for obtaining training vectors that adequately span both the acoustic signal space and the articulatory parameter space. Many different methods have been proposed to address this issue: interpolating from root shapes [272, 273] or random sampling [274], adaptive sampling [275], or training (either using neural networks [13, 276–279] or hidden Markov models [280]).

Many of these approaches use articulatory synthesizers to build the codebook (generally variations of Maeda's [281] or Mermelstein [282]'s articulatory models), but some of these only use the data obtained from acquisition (e.g. [278, 283]).

In the case of methods based on articulatory models, a quantity of data only limited by computational complexity can be used for training, but this data is dependent on the quality of the articulatory synthesizer.

In the case of methods that only use data from acquisition, the limited quantity of data is

generally a source of errors, since it does not span the whole acoustic space (and only a small subset of the articulatory space), and it is generally not sufficient for a statistically significant learning. We describe below several methods to build codebooks.

5.3.2.1 Step 1. Codebook construction

In [272], Larar and al. sample the articulatory space by a collection of "root" shapes representing the "most reasonable" articulatory regions for speech sounds. These root shapes were selected by closely matching the geometry and formant data for all vowels, by matching certain key features for consonant gestures, and three nasals were added. The articulatory space was the space of parameters of Mermelstein articulatory model [282]. From those 20 root shapes, the whole codebook was derived by sampling the straight lines from one root shape to another in the articulatory space to make a codebook containing about 10,000 shapes. These shapes were then clustered according to a measure of acoustic similarity based on LPC vectors.

In [273], a similar approach using "root shapes" is described, but with a different technique for interpolation. Instead of using a simple linear interpolation within the articulatory space, articulatory trajectories of diphones (vowel-consonant, consonant-consonant and vowel-vowel) are derived using an articulatory synthesizer. The exact procedure is unclear; we assume the transitions from one phone to another were constructed by interpolating the two target area functions, and the corresponding articulatory vectors were then derived. The articulatory vectors were added to the codebook only when the minimal cross-sectional area in the oral cavity changed more than 2 %. The resulting codebook was further optimised by regrouping articulatory vectors through linear approximation of the articulatory-to-acoustic relation.

5.3.2.1.1 Random sampling of the articulatory space In [274], Schroeter and al. used a random sampling method in order to have a much better acoustic coverage than in [272]. Samples were chosen randomly within wide margins of the articulatory models parameters, their acoustic images (represented as the first three formant frequencies) were computed, and a pruning method discarded points that were too similar (i.e., close articulatory vectors that give close acoustic images). Two different articulatory models were studied: Mermelstein [282] and Coker [266]. Thanks to this method, Schroeter and al. obtained a codebook that covered the acoustics of all the vowels and sonorants they studied, except one.

5.3.2.1.2 Neural networks Many different studies have attempted to use implicit codebooks obtained through neural neural networks or HMM training. The main interest of these methods is to accelerate the codebook look-up procedure. Different approaches have been adopted: training the neural networks on data from acquisition (corpus based), or training the network on data obtained from an articulatory model (articulatory model based), which are both problematic. Other approaches have trained the models on data sequences from acquisition corpus to avoid the problems of using articulatory models.

5.3.2.1.3 Adaptive sampling In [275], Ouni & Laprie use an adaptive sampling of the parameter space of Maeda's articulatory model. The articulatory space was explored recursively, and the subdivision is guided by a local sampling of the acoustic-to-articulatory relationship to check the local linearity of the relation: within an hypercube of the 7-D articulatory space, if the relationship is not linear enough according to a criterion involving every vertex of the hypercube, all 128 half-size sub-hypercubes will be explored. This method allows a complete description of the articulatory-to-acoustic relationship with a constant acoustic precision. For fast codebook look-up procedure, the 3D acoustic space of the three first formants frequencies is divided into cubes, and the articulatory hypercubes are ordered in the acoustic space according to the values of their images. Note that an hypercube can be associated to several different acoustic cubes. To generate actual inversion samples in a given hypercube, a particular sample is derived, and then the null space of the linear relation within this hypercube is explored using the simplex algorithm. Typical codebooks using this method, e.g. the one used in [284], contain more than 300,000 hypercubes, representing more than 4,000,000 pairs of articulatory and acoustic vectors. The precision of the codebook is almost sufficient to recover smooth trajectories directly from the codebook.

In several works, e.g. [273] and [285], the authors also use the local linearity of the relation to reduce the size of the codebook by regrouping samples that fall within the linearity domain of an articulatory vector. But these methods cannot guarantee, as [275] does, an homogeneous sampling of the whole articulatory space.

5.3.2.2 Step 2. Construction of initial trajectories

To construct an initial articulatory trajectory from the points obtained from the codebook lookup procedure, most studies penalize large "articulatory efforts", that is, fast changes in the vocal tract, and look for smoothly evolving articulatory trajectories under the constraint of matching a given sequence of speech spectra. This is conveniently done with dynamic programming, or even better with linear and non-linear filtering, like Kalman filtering or Ney's algorithm. Although diverse variations of these algorithms have been studied, almost all constraints used in the optimization process are derivatives of two kinds of components: **(i)** constraints on the acoustic vector in the codebook, to minimize the distance from the actual acoustic vector, and **(ii)** constraints on the articulatory parameters (almost all constraints of this kind are derivatives of articulatory trajectories efforts).

For instance, "muscle work criterion" was used for steady state segments in [11], as described in Section 5.2. The same kind of constraint was used in [267], but articulatory parameters all had the same weight, as opposed to the formulation in [11]. Pseudo-kinetic energy criteria have been used in [264], [275] and [286], where the two first studies also included a pseudo-potential. In most modern works, the articulatory criterion mixes static (potential energy) and dynamic (velocity and acceleration) features with varying weights. In some works, even the jerk (third order derivative of the position) has been used. Although these criteria are sometimes quite sophisticated, they usually are impaired by the lack of quantitative knowledge about the actual articulatory temporal behaviour.

More recently, other kinds of constraints have been investigated: in particular phonemic or phonetic constraints, that is, constraints applied according to the phonemic context, e.g.,

in [284] or [287]. Rough phonemic constraints are arguably already included within “root shapes interpolation” codebooks, although they are too strong, since only artificial trajectories from the codebook may be found in the inversed articulatory trajectories.

5.3.2.3 Step 3. Improvement of articulatory trajectories recovered

To improve the initial articulatory trajectories obtained from the previous step, which are generally rough and with some discontinuities, several methods have been proposed. In [274] an optimization scheme, using a gradient descent algorithm, optimise's the acoustic fit of the trajectory to obtain a spectrum as close as possible to the original. In this case, the ultimate goal was to use articulatory data for speech coding, so acoustic was the most important feature.

Most studies however, aim at recovering the most realistic articulatory trajectories possible, while conserving a good acoustic fit. In that prospect, methods of variational regularization have been used (e.g. [288]). In [275], the acoustic feature optimized was only the formant frequencies, whereas in most works the acoustic criterion is stronger. For example, Schroeter and Sondhi [286] use a complex cost function to evaluate the goodness of fit between the original and the synthetic signals that has four components, the first being the likelihood-ratio distance between the two LPC vectors corresponding to the original and the synthesized signals, the second comparing the energy of the two signals, the third comparing the time derivative of the glottal excitation; the fourth is a constraint on the articulatory vectors that penalizes high variations of the parameters.

The constraints applied in this optimization step are usually stronger than in the previous one, since the domain of available vectors is much greater, continuous (instead of the discretized space of the codebook values). The cost function minimized during this stage is usually the same as before, although additional components may be added.

5.4 Statistical data-based methods

When simultaneous acoustic and articulatory data is available, it is possible to based the inversion on statistical methods that ”learn” the quantitative association between the two types of data. Several types of mapping functions have been investigated, from linear estimators to more complex models based on HMM, as described in the following sections.

5.4.1 Linear estimation

With linear estimators, used e.g., in [17–19], unseen tongue data $\boldsymbol{X}$ is estimated from the acoustic input $\boldsymbol{Y}$ as:

$$\widetilde{\boldsymbol{X}} = \boldsymbol{T}_{XY} \cdot \boldsymbol{Y} \tag{5.1}$$

The estimator $\boldsymbol{T}_{XY}$ is defined from training data as

$$\boldsymbol{T}_{XY} = \boldsymbol{X} \cdot \boldsymbol{Y}^T \cdot (\boldsymbol{Y} \cdot \boldsymbol{Y}^T)^{-1} \tag{5.2}$$

The *N*-by-6 matrix $\boldsymbol{X}$ usually consists of EMA coil positions or articulatory parameters for each of the *N* time frames in the corpus. For the input data, LSP coefficients (described in Section 4.1.1) have been commonly used, so that $\boldsymbol{Y}$ is a *N*-by-*C* matrix, where each row contain the *C* LSP coefficients and the RMS amplitude (*C*=16 in all studies mentioned above).

The use of linear estimators has been most common in studies on the correlation between acoustics, vocal tract configuration and the face, and the results are further described in Section 5.5.

The articulatory-acoustic relation is however non-linear and a non-linear estimator, such as a neural network [289] or a relevance vector machine [109], may hence be more suitable. The differences between linear and these non-linear estimators are also described in Section 5.5.

5.4.2 Speech Inversion based on Hidden Markov Models

The HMM-Based Speech Production Model proposed in [290] allows the imposition of more elaborate constraints to the dynamic behaviour of the articulatory parameters that are estimated for given speech acoustics in a speech inversion setup. Its stochastic nature may also provide a formal way to cope with possible data measurement errors or imperfect assumptions. This model consists of phoneme-HMMs of articulatory parameters x and an articulatory-to-acoustic mapping that transforms the articulatory parameters into the speech spectrum y for each HMM state [290], as shown in Fig. 5.3. This mapping is approximated by the function $y = A_j x + b_j$ at state j. The covariance of the approximation error is σ_{w_j} while its mean is zero. Each phoneme HMM λ is defined as:

$$\lambda = \{\bar{x}_j, \sigma_{x_j}, \sigma_{w_j}, A_j, b_j, \bar{y}_j, \sigma_{y_j}, \alpha_{ij}\}, \text{for all j} \tag{5.3}$$

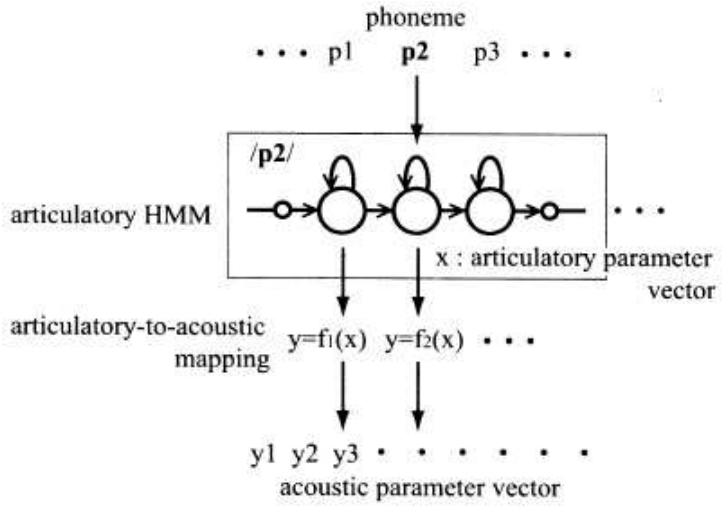


Figure 5.3: HMM-based speech production model [290]

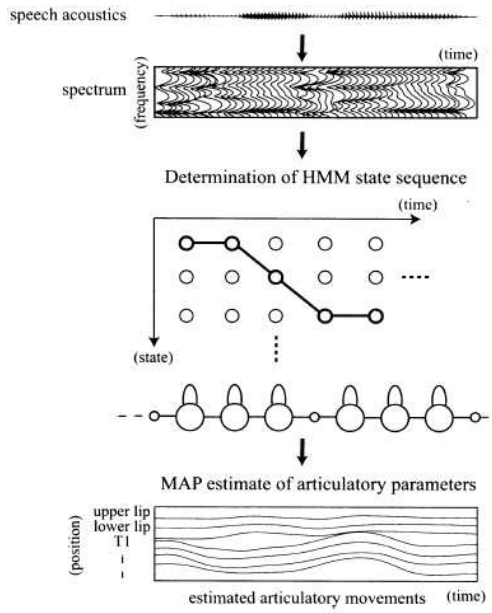


Figure 5.4: Procedure for acoustic-to-articulatory inverse mapping with HMMs [290]

The mean acoustic and articulatory parameter vectors at each state are $\bar{y}_j$ and $\bar{x}_j$ while the corresponding covariances are σ_{y_j} and σ_{x_j}. The transition probability from state i to j is α_{ij}.

Given a sequence of acoustic parameter vectors $\mathbf{y}$ and an HMM state sequence $\mathbf{q}$, the *estimation of the articulatory parameter vector sequence* $\mathbf{x}$ may be achieved by maximization of the a posteriori probability (MAP):

$$P(\mathbf{x}|\mathbf{y},\mathbf{q},\lambda) = \frac{P(\mathbf{y}|\mathbf{x},\mathbf{q},\lambda)P(\mathbf{x}|\mathbf{q},\lambda)}{P(\mathbf{y}|\mathbf{q},\lambda)} \propto P(\mathbf{y}|\mathbf{x},\mathbf{q},\lambda)P(\mathbf{x}|\mathbf{q},\lambda) \tag{5.4}$$

On a frame by frame basis, the desired parameter vector $\hat{x}$ is derived as a properly weighted sum of the mean articulatory vector at the current HMM state and the current vector x that satisfies the relationship $y = A_j x + b_j$. The weights are proportional to the relative reliability of the two summands:

$$\hat{x} = (\sigma_x^{-1} + A_j^T \sigma_w^{-1} A_j)^{-1}(\sigma_x^{-1}\bar{x} + A_j^T \sigma_w^{-1}(y - b_j)) \tag{5.5}$$

The estimation process is demonstrated in the Fig. 5.4. To limit abrupt changes of the estimates between subsequent frames the parameter vector may be enriched with the time derivatives and accelerations of the articulatory features. Then, with minor modifications, the described framework (MAP) may be applied to extract smooth articulatory parameters using the dynamic features as well [290].

The optimal state sequence may be determined using the Viterbi algorithm, as in the conventional acoustic HMMs [291]. The Gaussian observation probability distribution at each state with respect to the observed acoustic parameter vector is characterized by the following mean

and covariance:

$$\bar{y} = A_j \bar{x} + b_j \tag{5.6}$$

$$\sigma_y = A \sigma_x A^T + \sigma_w \tag{5.7}$$

Model training is performed using simultaneously obtained acoustic and articulatory data. Maximum likelihood estimation is achieved by means of the Expectation-Maximization algorithm. Firstly, the parameters $\bar{y}_j, \sigma_{y_j}$ and α_{ij} are obtained. Then, given the estimated probability $\gamma_t(j)$ of the acoustic parameter vector being at state j at the moment t as well as $\mathbf{x}$ and $\mathbf{y}$ we can determine the rest of the parameters by maximizing the a posteriori probability $P(\mathbf{x}|\mathbf{y}, \mathbf{q}, \lambda)$. Details are given in [290].

Reported experiments on data acquired by three Japanese males demonstrate improved performance of this HMM based speech inversion compared to two Codebook search methods, with and without a dynamic programming procedure [292]. The HMMs used were 3-state left to right diphone models. 873 diphone models plus an additional silence model were trained and used for inversion. The average RMS error was 1.73 mm while for the two Codebook search methods it was 2.35 mm and 2.95 mm, respectively.

5.4.2.1 Using Constrained HMMs

An alternative statistical approach for speech inversion is based on the so-called Constrained HMMs and is presented in [293]. The constrained hidden Markov models address the modeling of state dynamics by building some topology into the hidden state representation. The essential idea is to constrain the transition parameters of a conventional HMM so that the discrete-valued hidden state evolves in a structured way. The typical left-to-right constraints for the HMMs are a special case of the constrained state topologies which in general can be high-dimensional and allow omni-directional motion.

The definition of such an HMM involves the identification of each state of the hidden Markov chain as a spatial cell in a fictitious topology space. One has to select the dimensionality d for the space, the number of states M, the way these states will be packed (e.g., cubic) and the side length l of this packing. Dimensionality and packing define a vector-valued function $\mathbf{x}(m), \quad m = 1 \ldots M$ which gives the location of cell m in the packing. A constrained HMM in three dimensions is shown in Fig. 5.5. The most important is to choose a proper neighbourhood rule in the topology space, that is define which states belong to the same neighbourhood. The transition matrix of the HMM is then fixed so that it allows only transitions between neighbours. The non-zero transition probabilities may be set equally likely. The optimal state sequence is estimated using Viterbi decoding while for the state occupation probabilities the forward-backward algorithm is used. Constrained HMMs are trained in a similar way like the conventional HMMs. The main difference is that the transition probabilities are not updated by the Expectation Maximization algorithm.

Constrained HMMs may be applied to recover articulator movements from speech, as presented in detail in [293]. For training, simultaneous speech and articulatory data sequences are necessary. Speech is represented by a sequence of short-time spectral feature vectors (Line Spectral Frequencies). In the training phase, the idea is to learn the model parameters

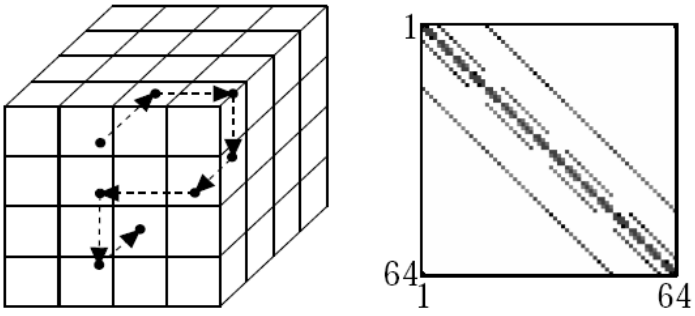


Figure 5.5: Constrained HMM for d=3, l=4, M=64 and cubic packing with an example state trajectory and the corresponding transition matrix [293]

so that connected paths through the state space can generate the speech training data with high likelihood. Model dimensionality and number of states are set using cross validation. After learning, one may infer a continuous state trajectory from an utterance by first generating a discrete state sequence m_t using Viterbi decoding and then interpolating smoothly between the positions $\mathbf{x}(m_t)$ of each state. A single linear fit is then performed between these state trajectories and the original movements of the articulators using the training data. For speech inversion, first the continuous state trajectory of the constrained HMM is inferred for the test utterance and then the single linear mapping is used to recover the exact articulator movements. Promising results are presented in [293] where the presented approach compares favourably to a Kalman based speech inversion technique using a global Linear Discrete State model.

5.5 Employing correlations between the face and vocal tract

Pure acoustic-to-articulatory inversion without constraints is theoretically impossible, due to the many-to-one mapping of several articulatory configurations to one speech spectrum and the fact that the system is under-determined with too few input parameters. The problem may to some extent be resolved by postulating that the speaker tries to minimize the energy and/or the articulatory distances or to maximize the smoothness of the movements. Another, more direct, method is to increase the number of input parameters, by adding what is already known about the articulation through other sources of information than the speech signal.

The most natural source of complementary information is to use data of the speaker's face. At the same time as the speech signal properties are defined by the configuration of the vocal tract, this configuration is reflected in the face, as the shaping of the vocal tract is to a large extent made by externally visible articulatory features, either directly, as the position of the jaw and the lip shape, or indirectly, as the tongue movements deform the skin, e.g. at the cheeks, through the muscle attachments.

Previous studies on the relation between the face, the vocal tract and the speech signal have used three sources of information for the face movements: automatic 3D tracking of infrared sensors glued to the face [17–19,128,289], manual tracking of coloured markers in video

images [294] or automatic tracking of facial features in video images [109].

Tracking of markers has the benefit that exact positions of the fleshpoints are given, for some systems even in three dimensions, but automatic analysis of an unmarked face is certainly more attractive for any application based on speech inversion, as the set-up is less complicated and intrusive for the speaker. In addition, it may also make use of information on shading or the visibility of the tongue tip, which is lost when tracking fleshpoints.

This section summarizes the results from previous studies on the relation between the face and vocal tract motion. The methods used to track the face and vocal tract are not described in any detail, as the methods as such have already been covered in section 2.

5.5.1 Does visual data help in the inversion?

Visual data does definitively improve the performance of the speech inversion, as indicated by the mean correlation coefficients of the previous studies in Fig. 5.6.

As suggested in Table 5.1, the studies differ in important aspects in the setup or method, reducing the validity of inter-study comparisons of correlation results. It is, on the other hand, possible to make comparisons within each study, with respect to the type of uni- or bi-modal input. Not only is the performance always better with the combined audiovisual input, these previous studies even suggest that information from the face is more important than the speech signal when trying to estimate the shape and position of the tongue. This conclusion holds regardless of corpus, method used to perform the regression and the type of visual input data.

Yehia et al. [17] (A in Fig. 5.6) explained the high correlation between the face and tongue movements by a functional coupling between the jaw and the tongue, i.e., that the two are moved together, except for the decoupling when the tongue, and especially the tip, is positioned independently of the jaw (typically for [l], which was observed to decrease the tongue motion

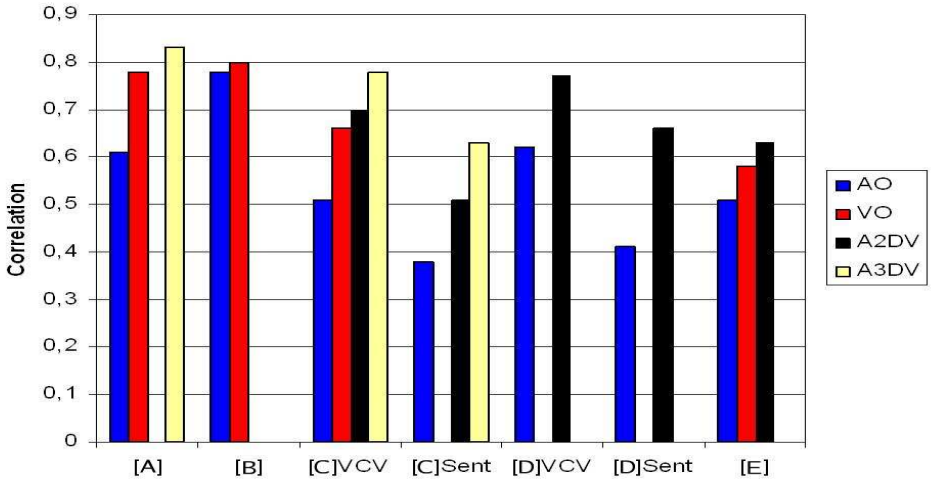


Figure 5.6: Mean correlation coefficients $\bar{\rho}^{AO}$ (audio input only), $\bar{\rho}^{VO}$ (visual only), $\bar{\rho}^{A2DV}$ (audio and 2D visual data) and $\bar{\rho}^{A3DV}$ (audio and 3D visual data) for the different studies listed in Table 5.1.

Corpus	Speakers/ Language	Facial data	Vocal tract description	Estimation method
		[A] Yehia et al., 1998 [17]		
2-5 sentences repeated	2 American, Japanese	3D tracking	7 EMA coils: tongue: 4, jaw: 1, lips: 2	linear
		[B] Jiang et al., 2002 [18]		
69 CVs 3 sentences repeated	4 American	3D tracking	5 EMA coils: tongue: 4, jaw: 1	linear
		[C] Engwall, 2005 [19]		
138 VCVs 178 sentences not repeated	1 Swedish	sparse 2D tracking	Articulatory parameters	linear
		[D] Engwall, 2006 [289]		
138 VCVs 178 sentences not repeated	1 Swedish	sparse 2D tracking	Articulatory parameters	ANN
		[E] Kjellström et al., 2006 [109]		
63 VCVs not repeated	1 Swedish	2D video of lips	4 EMA coils: tongue: 3, jaw: 1	RVM

Table 5.1: Previous studies on the correlation between the face, vocal tract and speech acoustics. The first column indicates the labels used in Fig. 5.6.

recovery). Similar results on the functional coupling were observed by Engwall & Beskow [19, 295].

Jiang et al. [18] (B in Fig. 5.6) on the other hand found that the lateral [l] was the best recovered manner of articulation for three of their subjects (but worst for the fourth one) when estimating the tongue from optical data, indicating that facial data can indeed capture tongue movements that are independent of the jaw. A word of caution is nevertheless called for, as both [17] and [18] used 4-5 repetitions of each syllable or phrase, which signifies that other productions of exactly the same utterance were included in the training material and there is hence a risk that the inversion is template-based rather than employing more general relations between features.

All the studies [19, 109, 128, 289, 295] were based on the KTH Qualisys-Movetrack database described in Section 2.3.4. The first four studies were based on motion capture data, while the last used video image analysis of the speaker's face.

Engwall [19] (C in Fig. 5.6) showed that while the full three-dimensional facial data improved the performance the most, a very sparse visual input, consisting of the five measures describing the horizontal position of the lip corner markers and the vertical position of the coils and the upper and lower lips and the jaw were enough to make a 40% improvement in the correlation

coefficients compared to the acoustic only input. Visual input describing a few selected features of the lip shape is hence a sufficient addition to improve the inversion results substantially. Study D in Fig. 5.6 that was based on the same data and conditions as C, but using a non-linear estimation method with neural networks, shows a similar improvement, but from a higher base-level.

The study by Kjellström et al. [109] (E in Fig. 5.6) further showed that the visual data need not be perfectly tracked by a dedicated optical motion tracking system in order to make a large improvement in the inversion results. Computer vision analysis of video images of the speaker's face gives a similar increase. The algorithm used was an Independent Component Analysis to describe the image features (c.f. Section 2.5.3) and a kernel-based relevance vector machine (RVM) to perform the regression between the vocal tract data and the face, acoustics or face plus acoustics. Two techniques were tested to combine the two sources of information; early fusion, in which the visual and acoustic data were merged before the regression, and late fusion, in which regression was performed on the two channels separately, before merging. The results for the A2DV case in Fig. 5.6 are for the late fusion, which was the better. The difference between the results for early and late fusion is discussed in the next section.

5.5.2 In what way does visual data help?

The contribution of facial data to the inversion can be analyzed in two different manners, either based on the regression results from facial data only or investigating the improvement when visual data is added to acoustic input.

5.5.2.1 Face-to-vocal tract estimation

In the study by Yehia et al. [17], the vertical jaw movement was very well reconstructed from the facial data (0.96 and 0.94 for the two subjects), and the level of correlation was almost identical for the vertical and horizontal movements of the tongue body and tip (0.7-0.76 for one subject, 0.9-0.91 for the other), except for the vertical movement of the tongue body for the second subject, where the correlation was somewhat lower (0.82). Yehia et al. performed no in-depth articulatory analysis of the correlation coefficients, but some additional information on the estimation of different parts of the vocal tract may be gained from the tongue motion traces of the estimated and measured tongue movement. The estimated tongue motion traces are more neutral than the real, especially for the horizontal movement. For the vertical movement, the estimation relies quite heavily on the strong relation between the jaw and tongue movements, in particular for the tongue tip. It thus seems that the more neutral the articulation is, and the more passive the tongue is, relative the jaw, the better will the estimation from facial motion capture data be.

Jiang et al. [18] divided the analysis of correlation results into categories depending on the place of articulation of the consonant to estimate, and the general finding over four speakers was that the groups of palatals, palatoalveolars, alveolars and dentals were the best estimated, even if the variation both on the levels and the rank between subjects was large. From the correlation coefficients it thus appears that articulations with a front lingual constriction are the

best reconstructed from facial data, which seems natural.

Engwall & Beskow [295] performed a similar investigation of the reconstruction of consonants with different places of articulation based both on correlation coefficients and on an articulatory analysis of tongue contours reconstructed from the real and estimated EMA coil positions. Instead of looking at the mean correlation coefficients for each consonant, they analyzed the reconstruction of different parts of the vocal tract. They found that the jaw position was almost perfectly estimated for all consonant groups; that the tongue body reconstruction was medium for all consonant groups except retroflexes, for which the estimation was poor; that the face provided little information for the tongue dorsum (i.e., the velar arching of the back part of the tongue), in general and especially for bilabials; that the tongue tip was quite well reconstructed for all consonants; and that the tongue advance was very badly estimated for all but bilabials.

When analyzing the tongue shapes, they concluded that the tongue tip position was well estimated for the majority of the alveolar stops and fricatives and retroflexes. On the other hand, the closure or constriction was not correctly estimated for velar stops and palatal and velar fricatives, as the vocal tract remained too open. For the fricatives, the constriction sometimes appeared, but too frontward. Moreover, the estimation of the lateral [l] was always unsuccessful, as the linear estimation from motion capture data assumes that a lowered jaw is accompanied by a low tongue tip, similar to the findings in [17]. The face did not give any information on the tongue position for the bilabials, since the tongue position is unconstrained when the closure is at the lips.

The face in itself clearly does not give sufficient information to recover the tongue contour for velars and bilabials, but it does provide important information on the front parts of the tongue for front articulations.

5.5.2.2 Audiovisual-to-vocal tract estimation

Engwall [19] made an extension of the work in [295] by investigating the combination of acoustic and sparse facial input data. The rank of the importance of the different facial measures differed between VCV words and sentences. For the VCV words, the lip corners were the most important (due to the importance of the lip rounding parameter in the corpus consisting of symmetric VCV words in [a, ɪ, ɔ] context), followed by the vertical positions of the upper and lower lip, and, close behind, the vertical movement of the jaw. For the sentences the upper-lower lip positions were the most beneficial, then the jaw and last the lip corners.

The facial measures provided the most information to recover the movements of the jaw and of the tongue tip raising, while the audio contributed the most to the horizontal position of the tongue tip. The largest increase gained by combining the two sources was for the front-back movement of the tongue body and the velar arching of the tongue. When grouping the VCV words depending on manner of consonant articulation, the largest increase when the facial data was added was for fricatives (64%) followed by nasals (53%). The increase for stops was significantly lower (29%) and for the approximant-tremulant group [l, j, h, r], the facial data actually decreased the performance (-12%), mainly due to the fact that the combination of a lowered jaw and a raised tongue tip for [l, r] goes against the general tendency in the corpus

that the movement of the jaw and tongue tip are positively correlated.

The five facial measures improved the estimation of the tongue tip position significantly and permitted to find e.g., alveolar closures. The estimation of manner of articulation was hence improved, even if the place of articulation was not always recovered for post-alveolars. The facial measures were unable to contribute to a better inversion of the tongue tip for articulations for which it was positioned very independently of the jaw, as for [l, r] and to the dorsum part of the tongue. It even occurred that the facial information contributed to a better estimation of the tongue tip position, but at the same time made the dorsum part correspondence worse. The facial information did nevertheless contribute more than (for VCVs) or as much as (for sentences) the acoustic signal to a successful recovery of the back part of the tongue, contrary to the commonly occurring statement that it is impossible to lip-read the back part of the tongue, at least for humans.

The results in [109], where automatic analysis of video images of the same subject and session was used, are similar. The audio-visual speech inversion outperformed both acoustic- and visual-to-articulatory inversion, and the visual data contributed more than the acoustic acoustic signal for all tongue coils, except for the horizontal position of the back-most tongue coil.

Compared to the results using motion capture of the face, the reconstruction of the jaw from the video images was not as perfect as from the 3D data, which is natural, since both the horizontal and vertical jaw movement is given almost directly by motion capture data, but must be estimated from the video. The horizontal movement is indicated only by changes in shading and the vertical movement needs to be estimated from the shape and size of the mouth opening rather than from an absolute position, since every image frame is centered on the lips. The vertical tongue tip position was estimated better from video images than from motion capture data, since the tongue tip is actually visible in some of the video images. For the remaining tongue coil coordinates, the estimation from video images was only marginally worse than that of the 3D motion capture, except for the back-most tongue coil. This is probably due to information given by markers on other parts of the face or the fact that the jaw position is almost perfectly estimated from the motion capture data. For all frames for which there is no independent tongue movement with respect to the jaw, the perfect recovery of the jaw will give a better estimation of the tongue.

The early fusion of audio and visual data was only marginally better than visual alone data, but late fusion resulted in a substantially higher correlation, which is in accordance with influential theories on human speech perception (e.g. [296]) stating that humans process information within each modality independently and then fuse the processed, rather than the raw, data. When analyzing the tongue shapes, it was found that the late fusion was better than early when the estimation from one of the modalities was close to the true shape and that the late fusion failed more gracefully than the early. The early fusion was on the other hand better when the estimation from both modalities failed or when one modality failed completely.

Engwall [289] compared linear and non-linear estimation of the tongue from acoustic and audiovisual data, using correlation coefficients, RMS errors and confusion matrices based on a articulatory classifier. The correlation coefficients and the RMS error indicate that the non-linear estimation (a neural network) performed better than the linear estimation, but also that the linear estimation gained more from adding visual features for VCV words, narrowing the gap

between the two methods. For sentences, the advantage of the non-linear estimation prevailed even when visual data was added.

To evaluate inversion results, an articulatory classifier was also proposed in [289]. It consists of prototype tongue shapes for each articulation, defined based on the frames labeled as belonging to the corresponding phoneme. The classifier was then employed to label every frame in the input data stream based on the articulatory correspondence with the prototypes, either regarding the entire tongue shape (vowels) or in the vicinity of the most constricted part of the vocal tract (for consonants). Confusion matrices could hence be defined to analyze the performance of the estimations, and they indicate that even if the overall levels of correctly classified articulations or places of articulation were similar for audiovisual input, the linear and non-linear estimations are quite different in terms of how they fail, both for acoustic only and audiovisual inversion. For the audiovisual inversion, misclassifications of the non-linear estimation tends to be closer in articulatory terms and the results were in particular better for alveolars and retroflexes compared to the linear estimation.

For the linear estimation, the main improvements when adding visual data were that the manner of articulation is better recovered for alveolars (e.g., [t]↔[s]) and the place of articulation better recovered for fricatives and back consonants.

For the non-linear estimation, the largest improvements were for [l, ɕ] and for the place of articulation for the palatovelars [ʃ, k], and a general redistribution of the misclassifications so that they were closer to the true articulatory category.

5.5.3 Summary & Discussion

In conclusion, visual data of the speaker's face is very important to recover the underlying configuration of the vocal tract, and it should hence be exploited in articulatory inversion, whenever it is possible. The studies summarized above indicate that the facial images are often even more important than the acoustic signal. It should be acknowledged, however, that the vocal tract configurations that were estimated in the above studies were represented by or reconstructed from a number of EMA coils on the tongue surface and on the jaw. It is thus possible, and even probable, that the acoustic signal is more important for other parts of the vocal tract, e.g., the configuration in the pharynx, which is not captured by the EMA measurements.

It could further be argued that most of the above studies have investigated the correlation on corpora containing speech material, e.g., VCV, CVC and CV, for which the correlation is very strong between the position of the jaw (easily estimated from facial data) and the configuration of the tongue. The two studies where sentences have been included in the corpus [19, 289] nevertheless show a similar improvement for sentences when visual data is added.

An additional possible caveat is the representation of the acoustic signal as line spectrum pairs. Though the LSP coefficients are considered to give a good representation of the acoustic spectrum and in particular the formants, it has yet to be shown that LSPs are the best possible representation for articulatory inversion. The representation might hence limit the contribution of the acoustic signal, and rather than claiming that visual data is *more* important than acoustic data, we wish to conclude that visual data is important in combination with the acoustic signal.

The one study that has investigated early and late fusion suggests that late fusion is better than early, as it gives more impact to a successful estimation from one of the modalities.

The most important contribution from the visual data is, quite naturally, for the estimation of the jaw position and the lip shape, but also for the tongue tip for front articulations. Even if the recovery of the back part of the tongue is less successful from visual data, it is often better than from the acoustic data only. The combined audiovisual inversion is, almost, always better than unimodal inversion. Problems may arise for articulations that do not correspond to the main relation between the important features of the face and the tongue, which may lead to a worse estimation if visual data is added. Noteworthy is, however, that such problems arise mainly for a linear estimation, while non-linear estimations, using, e.g., neural networks, achieve better results with audiovisual data even for these articulations. Non-linear estimation also seems to fail more gracefully than a linear. The audiovisual speech inversion should therefore ideally be performed with a non-linear estimation.

Chapter 6

Specification of fields investigated

This chapter is intended to present the scientific fields that will be investigated in the ASPI project. This presentation is brief because it refers to the technology inventory corresponding to chapters 1 to 5. Moreover, other deliverables, i.e. D2.1 about inversion methods, D4.1 about the design of the multimodal acquisition system and D5.1 give more details on the ongoing work.

6.1 Development of inversion methods

6.1.1 Tools for inversion

These tools are intended to prepare inversion by either providing data (automatic formant tracking) or adapting the analyzing model. The adaptation of the analyzing model mainly consists of modifying the geometrical dimensions of the articulatory model or the area function model. The overall fit between the analyzing model and the vocal tract of the speaker is assessed through the comparison of formant frequencies of synthesized speech and natural speech.

6.1.1.1 Automatic formant tracking

Although the objective is to perform audiovisual-to-articulatory inversion without the knowledge of formants, automatic formant tracking is important because it enables a very precise evaluation of inversion methods.

LORIA will improve formant tracking algorithms previously developed and make them available in order to build a formant database usable in the domain of acoustic-to-articulatory inversion.

The ICCS-NTUA group will contribute to the improvement of formant tracking by incorporating results from its on-going work on nonlinear speech modeling and tracking algorithms based on statistics and optimization. This work includes algorithms for detecting speech resonance modulations and estimating their parameters.

6.1.1.2 Speaker adaptation

Here, acoustical speaker adaptation designates the removal, or more precisely the attenuation, of the geometrical discrepancies between the speaker who uttered speech signal to be inverted and the analyzing model. An alternative is to perform speaker normalization by normalizing formant frequencies.

Speaker normalization will use MRI images that are now being acquired. There are already several normalization methods which often consist of adapting geometrical parameters of the articulatory model. The work will be about the evaluation of these methods and whether they need to be improved and how this can be achieved. The main issue will concern the recovery of the third dimension and its role in the speaker adaptation.

6.1.2 Improvement of the analyzing acoustic simulation

The objective is to guarantee that the analyzing model is able to approximate speech sounds correctly. Besides the synthesis itself, it is important to find out which are the main articulatory and acoustic characteristics that have to be exploited for inversion. The work of LTCI will be about the acoustic models of non-front sibilant fricatives.

In parallel, ICCS-NTUA will explore improvements to the analyzing model based on aeroacoustics and nonlinear speech production phenomena. This will be based on several experimental and theoretical evidences that such nonlinear aerodynamic phenomena occur during speech production and the development of signal processing systems that can approximate such phenomena.

A work carried out before ASPI about the MRI observation of non-front sibilant fricatives, /s and ʃ/, produced by 7 French speakers showed that there are two speaker-dependent strategies to make acoustic contrast between these consonants. It seems, moreover, that two simplified vocal-tract models with the noise-source location as the third parameter underlie the apparent inter-speaker variability, since these models can explain the observed spectral patterns of fricatives sounds produced by the those speakers. Finally, it is noted that these results would help us to formulate an appropriate vocal-tract model to be recovered from speech signal using an inverse method.

LTCI will thus address the issues described in Section 6.1.3.

6.1.3 Source of fricative sounds

Vocal vocal-tract configurations are recovered using an inverse method and we need to know how to control the generation of the fricative source in synthesis applications. The source generation requires a precise coordination between the articulation of the tongue (recovered by inversion) and the adduction/abduction of the larynx along a VCV sequence (not recovered by inversion), where V is a vowel and C is a fricative consonant. We shall carry out experimental studies observing the temporal relationships between open/closed phases of the glottis, the tongue movement, and airflow, which is flowed by an acoustic and airflow simulation studies.

The glottis and airflow can be observed by the instruments, respectively a photo-glottograph and a pneumotachograph, already existing in our lab. The tongue movements will be observed using a point-tracking device such as Aurora system. The knowledge gained from these experiments allows us to control the larynx for a given A-to-A inversed time-varying vocal-tract configuration along VCV sequences.

6.1.3.1 Tongue movements in VCV sequences

Using the Aurora system, LTCI shall study the contextual variation of tongue movements of fricatives in VCV contexts. In the literature, fricatives are said to have a high coarticulatory resistance, that is, they are resistant to the contextual influence of adjacent vowels. We feel however, from our experience that the front sibilant /s/ might have such a high resistance, but not non-front sibilant fricatives such as /ʃ/. We want to assess these observations by our own experiments. The result would be useful for the construction of experimental paradigm of inversion and for the evaluation of inverse methods.

6.1.3.2 Comparison of tongue contours derived from US imaging and X-ray films

Midsagittal x-ray data often describe complete tongue contours from the apex to the root. This is not the case in US imaging data, where the apex and tongue root regions are missing, although in the ASPI project the lacking tongue apex is recovered by tracking a magnetic sensor placed on the apex. The question here is whether the US derived tongue contours have sufficient information to define the whole tongue contours. This is not an unreasonable question, since we know from factor analysis that the tongue contour can be described, after subtracting the effect of the lower jaw position, by only two factor components with a high accuracy. If the net information content in tongue contours from X-rays and that from US imaging is identical or very close to each other, it must be a way to recover the root region of tongue contours from the US imaging data.

In addition, taking advantage of the superimposed point and US imaging data, it might be interesting to compare the flesh-point data and imaged tongue contour data in terms of kinematics. In the literature, it is said that the point data is superior in describing the kinematics than the imaged tongue contour data. It might be so in principle, but as far as we know, this claim is not validated by experiment. These two studies will be conducted by LTCI with a close collaboration with LORIA.

6.1.4 Inversion methods

The overall objective is to study the feasibility of inversion. We will thus address the following issues:

- Investigating the behavior of inversion, i.e. evaluating the shapes recovered from acoustic parameters according to the underlying analyzing model. Performing inversion with formant data enables the behaviour of the inversion to be clearly separated from other

factors, the properties of the spectral analysis for instance. The analyzing model can be represented globally by a codebook, or locally by estimating the Jacobian matrix of the articulatory to acoustic mapping at points visited by the inversion process.

- Elaboration of inversion methods that exploit spectral vectors like MFCC or LSP or alternatives (e.g., multiscale modulation-based) time-frequency speech representations and represent the articulatory to acoustic mapping by means of stochastic or neural models. Also, investigation of characteristics based on nonlinear speech modeling.

- Incorporation of constraints in order to reduce the under-determination of the inversion. These constraints can supplement the analyzing model by checking the phonetic, dynamical or anatomical consistency of inverse solutions. They can also provide articulatory information directly through the observation of the speaker's face.

6.1.4.1 Data-free formant-to-area mapping

The objective is to develop inverse mapping that obtains the shape parameters of a spectral model of the vocal tract from the measured formant frequencies, without training data. One reason for the focus on training data-free methods is the greater flexibility, because the switch between models or between different types of acoustic data does not request the compilation of a new code-book, which is time-consuming. Training data-free methods are therefore relevant even when the goal is inversion by table look-up, because the suitability of acoustic data or models for inverse mapping purposes can be evaluated before a final code-book is compiled.

The method that will be designed by ULB rests on the linearization of the relation between formant frequencies and the area function parameters; the model is spectral, that is, its eigenfrequencies are obtained directly from the tract parameters. No explicit model of the excitation or radiation is involved. Losses that have a major influence on the eigenfrequencies are inserted via geometric corrections to the vocal tract shape. The pseudo-inversion of the Jacobian matrix obtains the model parameter increments from the observed formant frequency increments. The parameter increments are added to the previous parameter values to obtain the present ones, after which the procedure is reiterated.

The Jacobian matrix is not square and its determinant is not zero. Its inverse is therefore replaced by its pseudo-inverse and the general solution is obtained by additional dynamic constraints with regard to position, speed, acceleration or jerk of the model parameters. Anatomical constraints are inserted by forcing the vocal tract parameters to evolve within bounds.

6.1.4.2 Data-free spectrum-to-area mapping

The objective is to develop an inverse mapping that obtains the shape parameters of a temporal or spectral model of the vocal tract from measured linear predictive or cepstral coefficients, without training data. The goal is to enable the spectral-to-shape inversion for arbitrary area function models, including losses, as well as voiced and fricative sources. Spectral cues are obtained algorithmically. They enable circumventing heuristics that obtain the formant frequencies. At present, data-free spectral-to-shape inversion may be considered to be an unsolved

problem, because existing methods either involve tract models that are over-idealized, or request prior table look-up.

The principle of the method proposed by ULB will rest on the linearization of the relation between spectral cues and the area function parameters; the models are spectral or temporal. Explicit models of the excitation or radiation are involved. Losses that have a major influence on the eigenfrequencies and bandwidths are taken into account. The pseudo-inversion of the Jacobian matrix obtains the model parameter increments from the observed spectral cue increments. The parameter increments are added to the previous parameter values to obtain the present values, after which the procedure is reiterated.

Similarly to the data-free formant-to-area mapping, anatomical constraints will be inserted.

6.1.4.3 Spectrum-to-articulatory and formant-to-articulatory mapping

The objective is to infer the parameters of articulatory models from acoustic data. The articulatory model is an articulatory-constrained area function model, or an articulatory model of the two-dimensional sagittal profile of a speaker's vocal tract. The purpose is to fix the area function by means of a small number of parameters that may have an anatomical or phonetic interpretation.

KTH, LORIA, ULB and ICCS-NTUA are involved in this task. They will investigate the inversion of mainly vowels and fricatives from spectral data (MFCC or similar spectral vectors), LSP (Line Spectral Pairs), or audio-visual features.

To achieve inversion, KTH, LORIA and ICCS-NTUA will investigate statistical learning methods.

KTH intends to investigate the influence of both different acoustic representations (e.g., MFCCs) and the learning method used. Previous results indicate that non-linear estimation methods are more valid than linear, and the work will hence be pursued to optimize estimations using e.g., multilayer perceptrons or mixed density networks.

LORIA will particularly study how the articulatory to acoustic mapping can be represented by a statistical learning and how it can be exploited for inversion.

In the above context, ICCS-NTUA will explore statistical learning approaches based on Hidden Markov Models.

6.1.5 Design and exploitation of constraints

The mapping from the articulatory space to the acoustical space is non-linear and many-to-one, and a challenge in audiovisual-to-articulatory inversion is thus to reduce the underdetermination of the problem through the incorporation of constraints. We will investigate physiological, phonetic and video constraints. Physiological constraints are put upon the physical characteristics of the articulators and can generally be expressed in the form of derivative terms that have to be minimized. Unlike other constraints they can be incorporated in the inversion process directly and issues that we will investigate are about their minimization and their con-

tributions to the inversion result.

KTH will continue to explore the constraints that geometric data of the face (3D motion capture of reflective markers on the face or video images) impose on the vocal tract configuration.

LORIA will investigate phonetic constraints derived from knowledge of the phonetic characteristics of speech sounds, which enables the derivation of reasonable domains in the space of articulatory parameters. Given the acoustical parameters of a speech sample (spectral vector or formant frequencies) an ”ideal” articulatory domain can be derived. The space of acoustical parameters is partitioned into sounds, using either speaker-specific data or generic. Then, to each articulatory vector can be associated a phonetic score varying with the distance to the ”ideal domain” associated with the corresponding sound. We will investigate whether phonetic constraints can be derived not only for vowels but also for consonants, and how these constraints can be expressed with formants and spectral vectors.

Video constraints derived from visible articulators come from the observation of the speaker's face and give information about the position of lips and the lower jaw. This corresponds to two or three articulatory parameters and this represents approximately one third of the global articulatory information. The constraints associated to visible articulators (when the speaker's face is visible) thus play a decisive role in the reduction of the under-determination of inversion. The usefulness of extracted visual features is expected to be two-fold. They can help in the establishment of quantitative articulatory constraints (off-line processing of audiovisual datasets while training the models) and can enhance auditory features for better interactive inversion (on-line processing of videos during the working phase of the system).

LORIA will investigate the exploitation constraints derived from 3D measures of the speaker's face.

ICCS-NTUA will also focus on the algorithmic and numerical frameworks that enable the incorporation of constraints, especially those of statistical pattern recognition/learning. Indeed, throughout the proposed project, new multimodal datasets will be gathered and, along with existing datasets, will be systematically processed. This will allow application of modern statistical inference techniques into the inversion problem. An important part of this research direction will deal with the optimal fusion of the features obtained by the different modalities.

6.1.6 Processing Video Images to Derive Constraints

In the design of the visual front-end of the audiovisual-to-articulatory inversion system ICCS-NTUA will address the following tasks:

- Active speaker's face detection and tracking: The appearance of the speaker's face contains significant information related to the configuration of articulators and thus the system must reliably locate and track it. Image pre-processing is usually required to suppress noise and spurious details and enhance the image/video. Geometric multiscale analysis (e.g., by means of morphological filters and/or partial differential equations-PDEs) is particularly effective in these tasks. At the same time, it respects the salient details of the image. Some effort will also be spent on extending methods to the time dimension. Efficient methods from computer vision will be utilized next for detecting the speaker's face

and tracking.

- Facial model fitting and visual features extraction: After the speaker's face has been detected, the position and dynamics of the visible articulators (mainly lips and jaw) must be accurately captured. Face modeling has attracted significant research interest in the scientific community and a multitude of methods have been proposed. We plan to test the performance of both advanced active shape models as well as active appearance models. The parameters describing the fitted face model, after appropriate dimensionality reduction, will be used as articulatory visual features. In addition to extracting features from the lips and jaw area, some effort will be spent on detecting and tracking the vocal tract shape on data that provide such information.

6.2 Design, acquisition and processing of articulatory data

6.2.1 Defining acquisition protocols

KTH and LORIA will define acquisition protocols for all the imaging modalities used in the project. The repeatability and, the quality of 3D and temporal data will receive a particular attention. Ultrasound imaging will be used in order to evaluate the impact of the acquisition conditions (supine or sitting position, real or silent production. . .).

6.2.2 Acquisition of data

The content of the articulatory database will be defined by KTH and LORIA. The first objective is to enable the development of a dynamic articulatory model as complete as possible for two Swedish and two French speakers (if possible more speakers will be involved). The second objective is to record an amount of dynamic data that enables the training and the evaluation of inversion methods.

6.2.3 Exploitation and Processing of Databases

Recovery of geometric information regarding the articulators is both intrinsically interesting and required in the ASPI project. This needs to be done both in existing X-ray data as well as in images acquired using other modalities such as US or MRI. Efficient image processing and computer vision techniques for enhancement, segmentation and tracking of medical images need to be utilized or further developed. In this area, ICCS-NTUA will address the following tasks:

- *Part I: Image Pre-Processing and Enhancement*:

  De-noising,

  Contrast Enhancement, Interpolation, and Multiscale Image Simplification

Possible Methods:
Multiscale Morphological Operators,
Vision PDEs, Variational and Level set methods,
Multiband filtering

- *Part II: Segmentation and Boundary Tracking*:

  Extract the shape of internal articulators from images.

  Possible Methods: Active Contours, Morphological, and PDE methods.

  Incorporation of Prior Knowledge.

  Extensions: Boundary tracking in videos.

6.3 Multimodal acquisition technology

Beside more technical aspects, two issues will be addressed to enable the development of the multimodal acquisition system.

6.3.1 Tongue tracking on ultrasound images

Recovery of geometric information regarding the tongue in US images is very important for the overall success of the project. Recent image processing techniques for segmentation and tracking of medical images will be utilized. It must be noted that some techniques used for curve detection in X-ray images will also be used for curve tracking in other modalities as in US images or MRI images.

LORIA will address this problem and there will be a cooperation with ICCS-NTUA on this topic which is close to the processing of existing X-ray databases.

6.3.2 Fusing tongue tracking and other modalities to recover the complete shape of the tongue including the apex

As the various image/sensor devices are not referenced the same way, an important and difficult task before combining the image/sensor modalities is to register all the data in order to express them in a reference spatial and temporal frame.

LORIA will address the following aspects:

- the stereovision/US registration. We plan to register the two modalities by acquiring a sequence of the head with the US transducer positioned under the chin. 3D markers will be added on the handle in order to get the probe position in the video frame.

- the sensor/stereovision registration. We plan to use several magnetic sensors (at least four): two of them will be positioned on the tongue for articulatory modeling. The others two will be positioned on the head. Registration can be achieved from these external sensors whose position can be estimated both by the stereovision and the sensor system. Another possible solution is to point out the markers that are drawn on the speaker's head with a sensor coil.

- stereovision/MRI registration. Due to the time needed for MRI acquisition, it is likely that we will only acquire a thick sagittal slice of the head. As the skin can be easily extracted from MRI, registration will be achieved using an Iterative Closest Point algorithm in order to minimize the distance between the set of markers drawn on the head and the surface of the skin extracted from the MRI data.

- speech alignment: dynamic programming will be used to align speech segment as described in [69, 70] for tongue reconstruction from ultrasound images.

The temporal synchronization of images, sound and magnetic sensors, which is a central issue in the fusion, will exploit events that occur in several modalities and hardware solutions.

Bibliography

[1] M. M. Sondhi, "Estimation of vocal tract areas: the need for acoustical measurements," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 268–272, 1979.

[2] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *J. Acoust. Soc. Am*, vol. 41, pp. 1283–1294, 1967.

[3] H. Wakita, "Estimation of vocal tract shapes from acoustical analysis of the speech wave: the state of the art," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 281–285, 1979.

[4] J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer-Verlag, Berlin Heidelberg New York, 1976.

[5] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique," *Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1535–1555, May 1978.

[6] M. Rahim, *Artificial neural network for speech analysis/synthesis*, Chapman and Hall, London, 1994.

[7] S. Ouni and Y. Laprie, "Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 444–460, 2005.

[8] L.-J. Boë, P. Perrier, and G. Bailly, "The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion," *Journal of Phonetics*, vol. 20, pp. 27–38, 1992.

[9] F. Charpentier, "Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic non-linearities," *Speech Communication*, vol. 3, pp. 291–308, 1984.

[10] M. Mrayati and R. Carré, "The acoustic-area function inversion problem and the distinctive region model," in *Signal processing VI*, J. Vandewalle, R. Boîte, M. Moonen, and A. Osterlinck, Eds. 1992, pp. 155–158, Elsevier Publishers.

[11] V.N. Sorokin, A.S. Leonov, and A.V. Trushkin, "Estimation of stability and accuracy of inverse problem solution for the vocal tract," *Speech Communication*, vol. 30, pp. 55–74, 2000.

[12] H. Yehia and F. Itakura, "Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal fourier analysis," in *Proc. IEEE Int. Conf. ASSP*, 1994, vol. 0-7803-1775-0/94.

[13] A. Soquet, M. Saerens, and P. Jospa, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model: further results," in *Artificial Neural Networks*, O. Simula T. Kohonen, K. Mäkisara and J. Kangas, Eds., pp. 371–376. Elsevier, 1991.

[14] J. Flanagan, K. Ishizaka, and K. Shipley, "Signal models for low bit-rate coding speech," *Journal of the Acoustical Society of America*, vol. 68, no. 3, pp. 780–791, March 1980.

[15] V. N. Sorokin, "Determination of vocal tract shape for vowels," *Speech Communication*, vol. 11, pp. 71–85, 1992.

[16] S. Tasko and J. R. Westbury, "Speed-curvature relations for speech-related articulatory movement," *J. Phonetics*, vol. 32, pp. 65–80, 2004.

[17] Hani Yehia, Philip E. Rubin, and Eric Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1–2, pp. 23–44, 1998.

[18] J. Jiang, J. Alwan, P. Keating, and L. Auer, E.and Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174–1188, 2002.

[19] O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in *Proceedings of Interspeech*, 2005, pp. 3205–3208.

[20] H. Yehia and F. Itakura, "A method to combine acoustic and morphological constraints in the speech production inverse problem," *Speech Communication*, vol. 18, no. 2, pp. 151–174, 1996.

[21] H. Wakita and A. Gray, "Numerical determination of the lip impedance and vocal tract area functions," *IEEE Transactions on Acoustics, Speech, Sig. Proc.*, vol. 23, 6, pp. 574–580, 1975.

[22] K. Shirai and M. Honda, "Estimation of articulatory motion from speech waves and its application for automatic recognition," in *Spoken Language Generation and Understanding*, J. C. Simon, Ed. 1980, pp. 87–99, D. Reidel Publ. Co.

[23] P. Ladefoged, R. Harshman, L. Goldstein, and L. Rice, "Generating vocal tract shapes from formant frequencies," *J. Acoust. Soc. Am*, vol. 64, 4, pp. 1027–1035, 1978.

[24] M. M. Sondhi and J. R. Resnick, "The inverse problem for the vocal tract: numerical methods, acoustical experiments, and speech synthesis," *J. Acoust. Soc. Am*, vol. 73, 3, pp. 985–1002, 1983.

[25] Z. Yu, "A method to determine the area function of speech based on perturbation theory," *Speech Transmission Laboratory, QPSR*, vol. 4, pp. 77–95, 1993.

[26] D. Rossiter, D. M. Howard, and M. Downes, "A real-time LPC-based vocal tract area display for voice development," *J. Voice*, vol. 8, 4, pp. 314–319, 1994.

[27] I. Kamal, *Acoustic reflectometry of the nose and pharynx*, Brown Walker Press, Boca Raton, Florida, USA, 2004.

[28] I. Gath and E. Yair, "Analysis of vocal tract parameters in Parkinsonian speech," *J. Acoust. Soc. Am*, vol. 84, 5, pp. 1628–1634, 1988.

[29] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am*, vol. 41, 2, pp. 1002–1010, 1967.

[30] Z. Yu and P. C. Ching, "Determination of vocal-tract shapes from formant frequencies based on perturbation theory and interpolation method," in *IEEE-ICASSP, typed manuscript*, 1996.

[31] H. Fujisaki, S. Obata, and R. Tazaki, "Estimation of vocal tract area function from poles of its transfer function," in *Annual Report of the Engineering Research Institute, Faculty of Engineering, University of Tokyo*, 1971, vol. 30, pp. 81–94.

[32] G. Bailly, C. Abry, R. Laboissièrre, P. Perrier, and J.-L. Schwartz, "Inversion and speech recognition," in *Signal processing VI: Theories and Applications*, J. Vandewalle, R. Boite, M. Mooner, and A. Osterlinck, Eds., Brussels, Belgium, August 1992, vol. 1, pp. 159–164, Elsevier.

[33] T. Nakajima, H. Omura, and S. Ishizaka, "Estimation of the vocal tract area functions by adaptive inverse filtering methods and identification of articulatory model," in *Proc. Speech Communication Seminar, Stockholm*, G. Fant, Ed., New York, 1975, pp. 11–21, Wiley.

[34] M. G. Rahim and C. C. Goodyear, "Estimation of vocal tract filter parameters using a neural net," *Speech Communication*, vol. 9, pp. 49–55, 1990.

[35] S. K. Gupta and J. Schroeter, "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2517–2530, Nov 1993.

[36] P. Meyer, R. Wilhelms, and H. W. Strube, "A quasi-articulatory speech synthesizer for german running in real time," *J. Acoust. Soc. Am*, vol. 86, 2, pp. 523–539, 1989.

[37] K. Shirai, "Estimation and generation of the articulatory movement using neural networks," *Speech Communication*, vol. 13, pp. 45–51, 1993.

[38] K. Shirai and T. Kobayashi, "Estimating articulatory movement from speech wave," *Speech Communication*, vol. 5, pp. 159–170, 1986.

[39] K. Shirai and M. Honda, "Morphological filtering for image enhancement and detection," in *Spoken Language Generation and Understanding*, J. C. Simon, Ed., pp. 87–99. D. Reidel Publ. Co., 1980.

[40] P. Prado, E. Shiva, and D. Childers, "Optimization of acoustic-to-articulatory mapping," in *Proc. ICASSP*, 1992, pp. 33–36.

[41] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *J. Acoust. Soc. Am*, vol. 74, pp. 1145–1154, 1983.

[42] G. O. Russell, *The vowel, its psychological mechanism, as shown by x-ray*, Ohio State University Press, Columbus, OH USA, 1928.

[43] S. Dart, "A bibliography of x-ray studies of speech," *UCLA Phonetics Laboratory Group*, vol. 66, 1987.

[44] G. Fant, *Acoustic Theory of Speech Production*, Mouton, the Hague, Netherlands, 1960.

[45] G. Fant, "Formants and cavities," in *Proc of ICPhS'65*, 1965, pp. 120–140.

[46] G. Fant, "Feature analysis of Swedish vowels - a revisit," *STL-QPSR*, vol. 2-3, pp. 1–19, 1983.

[47] J. Stark, C. Ericsdotter, P. Branderud, J. Sundberg, H.-J. Lundberg, and J. Lander, "The apex model as a tool in the specification of speaker-specific articulatory behavior," in *Proc of ICPhS*, 1999, pp. 2279–2282.

[48] S. Kiritani, K. Itoh, and O. Fujimura, "Tongue-pellet tracking by a computer controlled x-ray microbeam system," *Journal of the Acoustic Society of America*, vol. 48, pp. 1516–1520, 1975.

[49] O. Fujimura, "Recording and interpreting articulatory data – microbeam and other methods," in *Proc of ICPhS91*, 1991, vol. 3, pp. 120–124.

[50] P. Badin, E. Baricchi, and A. Vilain, "Determining tongue articulation: from discrete fleshpoints to continuous shadow," in *Proc of Eurospeech97*, 1997, vol. 1, pp. 47–50.

[51] R. Beaudoin and R. McGowan, "Principal component analysis of x-ray microbeam data for articulatory recovery," in *Proc of the 5^{th} Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, 2000, pp. 225–228.

[52] J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Matthies, I. Garabieta, and M. T. T. Jackson, "Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements," *J. Acoust. Soc. Am*, vol. 92, pp. 3078–3096, 1992.

[53] P. Hoole, "Methodological considerations in the use of electromagnetic articulography in phonetic research," Tech. Rep. Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation, Universität München, 1993.

[54] P. Branderud, "Movetrack - a movement tracking system," in *the French-Swedish Symposium on Speech*, Grenoble, France, 1985, pp. 113–122.

[55] A. Zierdt, P. Hoole, M. Honda, T. Kaburagi, and H. Tillman, "Extracting tongues from moving heads," in *Proc of the 5^{th} Speech Production Seminar: Models and data*, 2000, pp. 313–316.

[56] L. Fitzpatrick and A. Chasaide, "Human speaker nomograms using EMA data," in *Proc of ICPhS*, 1999, vol. 3, pp. 2021–2024.

[57] N. Nguyen-Trong, P. Hoole, and A. Marchal, "Articulatory-acoustic correlation in the production of fricatives," in *Proc of ICPhS91*, 1991, pp. 1:18–21.

[58] O. Engwall, "Combining MRI, EMA & EPG in a three-dimensional tongue model," *Speech Communication*, vol. 41, no. 2–3, pp. 303–329, 2003.

[59] C. Kelsey, F. Minifie, and T. Hixon, "Applications of ultrasound in speech research," *J Speech and Hearing Research*, vol. 12, pp. 564–575, 1969.

[60] E. Keller and D. Ostry, "Computerized measurement of tongue dorsum movements with pulsed echo ultrasound," *J Acoust Soc Am*, vol. 73, pp. 1309–1315, 1983.

[61] F. Minifie, C. Kelsey, and J. Zagzebski, "Ultrasonic scans of the dorsal surface of the tongue," *J Acoust Soc Am*, vol. 49, pp. 1857–1860, 1971.

[62] B. Sonies, T. Shawker, T. Hall, L. Gerber, and S. Leighton, "Ultrasonic visualization of tongue motion during speech," *J Acoust Soc Am*, vol. 70, pp. 683–686, 1981.

[63] A. Parush, D. Ostry, and K. Munhall, "A kinematic study of lingual coarticulation in vcv sequences," *Journal of the Acoustic Society of America*, vol. 74, pp. 1115–1125, 1983.

[64] E. Keller, "Ultrasound measurements of tongue dorsum movements in articulatory speech impairments," in *Phonetic Approaches to Speech Production in Aphasia and Related Disorders*, J.H. Ryalls, Ed., pp. 93–112. College-Hill Press, San Diego, CA, 1987.

[65] A. Lundberg and M. Stone, "Three-dimensional tongue surface shapeswreconstruction: Practical consideration for ultrasound data," *Journal of the Acoustic Society of America*, vol. 106, pp. 2858–2867, 1999.

[66] M. Stone, B. C. Sonies, T. H. Shawker, G. Weiss, and L. Nadel, "Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system," *Journal of Phonetics*, vol. 11, pp. 207–218, 1983.

[67] M. Stone and E. David, "A head and transducer support system for making ultrasound images of tongue/jaw movement," *J. Acoust. Soc. Am*, vol. 98, pp. 3107–3112, 1995.

[68] D. H. Whalen, K. Iskarous, M. T. Tiede, D. Ostry, H. Lehnert-LeHoullier, and D. Hailey, "The haskins optically-corrected ultrasound system (HOCUS)," *Journal of Speech, Language, and Hearing Research*, vol. 48, pp. 543–553, 2005.

[69] Y. Akgul, C. Kambhamettu, and M. Stone, "A task-specific contour tracker for ultrasound," in *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2000, pp. 135–142, http://citeseer.ist.psu.edu/613227.html.

[70] M. Li, C. Kambhamettu, and M. Stone, "Edgetrak, a program for band-edge extraction and its applications," in *The sixth IASTED International Conference on Computers, Graphics and Imaging*, 2003, http://www.speech.umaryland.edu/Publications/Minf.

[71] M. Rokkaku, K. Hashimoto, S. Imaizumi, S. Nimi, and S. Kirtani, "Measurements of the three-dimensional shape of the vocal tract based on the magnetic resonance imaging technique," *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, vol. 20, pp. 47–54, 1986.

[72] T. Baer, J.C. Gore, S. Boyce, and P.W. Nye, "Application of MRI to the analysis of speech production," *Magnetic Resonance Imaging*, vol. 5, pp. 1–7, 1987.

[73] T. Baer, J.C. Gore, L.W. Gracco, and P.W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustic Society of America*, vol. 90, pp. 799–828, 1991.

[74] O. Engwall, "Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation," in *Speech production: Models, Phonetic Processes and Techniques*, J. Harrington and M. Tabain, Eds., pp. 301–314. Psychology Press, New York, 2006.

[75] P. Badin, G. Bailly, L. Revéret, M. Baciu, C. Segebarth, and C. Savariaux, "Three-dimensional articulatory modelling of tongue, lips and face, based on MRI and video images," *Journal of Phonetics*, vol. 30, no. 3, pp. 533–553, 2002.

[76] M. Tiede, S. Masaki, and E. Vatikiotis-Bateson, "Contrasts in speech articulation observed in sitting and supine condition," in *The 5th Speech Production Seminar: Model and data*, 2000, pp. 25–28.

[77] A-K. Foldvik, U. Kristiansen, and J. Kvaerness, "A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI)," in *Proc of Eurospeech93*, 1993, pp. 557–558.

[78] M. Mohammad, E. Moore, J. Carter, C. Shadle, and S. Gunn, "Using MRI to image the moving vocal tract during speech," in *Proc of Eurospeech97*, 1997, pp. 2027–2030.

[79] C. Shadle, M. Mohammad, P. Jackson, and J. Carter, "Multi-planar dynamic magnetic resonance imaging: New tools for speech research," in *Proc of ICPhS*, 1999, pp. 623–626.

[80] S. Masaki, M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto Y. Nakamura, and N. Ninomia, "MRI-based speech production study using a synchronized sampling method," *Journal of Acoustical Society of Japan*, vol. 20, pp. 375–379, 1999.

[81] M. Stone, D. Dick, A. Douglas, E. Davis, and C. Ozturk, "Modelling the internal tongue using principal strains," in *Proc of the* 5^{th} *Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, 2000, pp. 133–136.

[82] D. Demolin, M. George, V. Lecuit, T. Metens, A. Soquet, and H. Raeymaekers, "Coarticulation and articulatory compensations studied by dynamic MRI," in *Proc of Eurospeech*, 1997, pp. 43–46.

[83] K. Mády, R. Sader, A. Zimmermann, P. Hoole, A. Beer, H.F. Zeilhofer, and C. Hanning, "Use of real-time MRI in assessment of consonant articulation before and after tongue

surgery and tongue reconstruction," in *Proc of the 4th International speech motor conference: speech motor control in normal and disordered speech*, 2001, pp. 142–145.

[84] S. Narayanan, KS. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 1771–1776, 2004.

[85] A. Katsamanis, G. Papandreou, V. Pitsikalis, and P. Maragos, "Multimodal fusion by adaptive compensation for feature uncertainty with application to audiovisual speech recognition," in *EUSIPCO*, 2006.

[86] R. Koch, M. Gross, F. Carls, D. Büren, G. Fankhauser, and Y. Parish, "Simulating facial surgery using finite element models," *Computer Graphics*, vol. 30, no. Annual Conference Series, pp. 421–428, 1996.

[87] E. Okada, "Three-dimensional facial simulations and measurements: Changes of facial contour and unitsassociated with facial expression," *Journal of Craniofacial Surgery*, pp. 167–174, 2001.

[88] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH '99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 1999, pp. 187–194, ACM Press/Addison-Wesley Publishing Co.

[89] W-S. Lee and N. Magnenat-Thalmann, "Fast head modeling for animation.," *Image Vision Comput.*, vol. 18, no. 4, pp. 355–364, 2000.

[90] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin, "Synthesizing realistic facial expressions from photographs," *Computer Graphics*, vol. 32, no. Annual Conference Series, pp. 75–84, 1998.

[91] N. A. Borghese and S. Ferrari, "A portable modular system for automatic acquisition of 3-D objects," *IEEE Trans. on Instr. and Meas.*, vol. 49, no. 5, pp. 1128–1136, Oct. 2000.

[92] D. DeCarlo, D. Metaxas, and M. Stone, "An anthropometric face model using variational techniques," in *SIGGRAPH '98: Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, New York, NY, USA, 1998, pp. 67–74, ACM Press.

[93] Pascal Fua, "Regularized bundle-adjustment to model heads from image sequences without calibration data," *International Journal of Computer Vision*, vol. 38, no. 2, pp. 153–171, 2000.

[94] Y. Shan, Z. Liu, and Z. Zhang, "Model-based bundle adjustment with application to face modeling.," in *ICCV*, 2001, pp. 644–651.

[95] S. Marschner, B. Guenter, and S. Raghupathy, "Modeling and rendering for realistic facial animation," in *Proceedings of the Eurographics Workshop on Rendering Techniques 2000*, London, UK, 2000, pp. 231–242, Springer-Verlag.

[96] Cyberware, "Synthesizing realistic facial expressions from photographs," *Head and Face Color 3D Scanner Model 3030*.

[97] M. Proesmans and L. V. Gool, "Reading between the lines: a method for extracting dynamic 3D with texture," in *Proceedings of the ACM symposium on Virtual reality software and technology*, NY, USA, 1997, pp. 95–102, ACM Press.

[98] R. Sitnik and M. Kujawinska, "Opto-numerical methods of data acquisition for computer graphics and animation systems," in *Proc. SPIE Vol. 3958, p. 36-43, Three-Dimensional Image Capture and Applications III, Brian D. Corner; Joseph H. Nurre; Eds.*, B. D. Corner and J. H. Nurre, Eds., Mar. 2000, pp. 36–43.

[99] S. Romdhani and T. Vetter, "Efficient, robust and accurate fitting of a 3D morphable model," in *Proc. Int'l Conf. on Comp. Vision*, 2003, pp. 59–66.

[100] H. Shing and L. Yin, "Constructing a 3D individualized head model from two orthogonal views.," *The Visual Computer*, vol. 12, no. 5, pp. 254–266, 1996.

[101] Y. Matsumoto, K. Fujimura, and T. Kitamura, "Cybermodeler: A compact 3D scanner based on monoscopic camera.," in *Three-Dimensional Image Capture and Applications*, 1999, pp. 2–10.

[102] J.Y. Zheng, "Acquiring 3-d models from sequences of contours," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 163–178, 1994.

[103] R. Kaucic and A. Blake, "Accurate, real-time, unadorned lip tracking," in *ICCV*, 1998, pp. 370–375.

[104] R. Seymour, J. Ming, and D. Stewart, "A new posterior based audio-visual integration method for robust speech recognition," in *Interspeech*, 2005, pp. 1229–1232.

[105] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *ICASSP*, 1994, pp. 669–672.

[106] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[107] I. Matthews, T. F. Cootes, J. A. Banghamand S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *PAMI*, vol. 24, no. 2, pp. 198–213, 2002.

[108] I. Shdaifat and R-R. Grigat, "A system for audio-visual speech recognition," in *Interspeech*, 2005, pp. 1221–1224.

[109] H. Kjellström, O. Engwall, and O. Bälter, "Reconstructing tongue movements from audio and video," in *Interspeech*, Pittsburgh, Sep 2006, pp. 2238–2241.

[110] K. Saenko, K. Livescu, M. Siracusa, K.Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *ICCV*, 2005, pp. 1424–1431.

[111] F. Elisei, M. Odisio, G. Bailly, and P. Badin, "Creating and controlling video-realistic talking heads," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'01)*, Aalborg, Denmark, 2001, pp. 90–97.

[112] G. Kalberer and L. Van Gool, "Face animation based on observed 3D speech dynamics," in *Proceedings of Computer Animation Conference*, 2001, pp. 20–27.

[113] Robert V., Wrobel-Dautcourt B., Laprie Y., and Bonneau A., "Strategies of labial coarticulation," in *Interspeech, Lisboa*, Sept. 2005, pp. 1021–1024.

[114] B. Wrobel-Dautcourt, M. O. Berger, B. Potard, Y. Laprie, and S. Ouni, "A low cost stereovision based system for acquisition of visible articulatory data," in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'05)*, Vancouver, 2005, pp. 145–150.

[115] Kanade and Okutomi, "A stereo matching algorithm with adaptive window: theory and experiments," *IEEE Transactions on PAMI*, vol. 16, no. 9, pp. 920–932, 1994.

[116] J-P. Pons, R. Keriven, O. Faugeras, and G. Hermosillo, "Variational stereovision and 3D scene flow estimation with statistical similarity measures," in *ICCV*, 2003, pp. 597–602.

[117] Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen, "Rapid modeling of animated faces from video images," in *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, New York, NY, USA, 2000, pp. 475–476, ACM Press.

[118] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999.

[119] P. Fua and C. Miccio, "Animated heads from ordinary images: A least squares approach," *Computer Vision and Image Understanding*, vol. 75, no. 3, pp. 247–259, 1999.

[120] L. Zhang, L. Curless, and S. Seitz, "Spacetime stereo: shape recovery for dynamic scenes," in *IEEE Int Conf on Computer Vision and Pattern Recognition*, 2003, pp. 367–374.

[121] Li Zhang, Noah Snavely, Brian Curless, and Steven M. Seitz, "Spacetime faces: High-resolution capture for modeling and animation," in *ACM Annual Conference on Computer Graphics (Los Angeles)*, 2004, pp. 548–558.

[122] K. G. Munhall, E. Vatikiotis-Bateson, , and Y. Tohkura, "X-ray film database for speech research," *J. Acoust. Soc. Am*, vol. 98, pp. 1222–1224, 1995.

[123] C. Rochette, *Les groupes de consonnes en français*, Les Presses de l'Université Laval, Québec, Canada, 1973.

[124] J. S. Perkell, *Physiology of speech production: results and implications of a quantitative cineradiographic study*, MIT Press, Cambilidge, Mass. USA, 1969.

[125] A. Arnal, P. Badin, G. Brock, P.-Y. Connan, E. Florig, N. Perez, P. Perrier, P. Simon, R. Sock, L. Varin, B. Vaxelaire, and J.-P. Zerling, "Une base de données cinéradiographiques du français," in *XXIIIèmes Journées d'Etude sur la Parole*, 2000, pp. 425–428.

[126] J. R. Westbury, G. Turner, and J. Dembovski, *X-ray microbeam speech production database user's handbook*, Waisman Center, University of Wisconsin, Madison, WI USA, 1994.

[127] A. Wrench and W. Hardcastle, "A multichannel articulatory speech database and its application for automatic speech recognition," in *5th International Seminar on Speech Production*, 2000, pp. 305–308.

[128] J. Beskow, O. Engwall, and B. Granström, "Resynthesis of facial and intraoral motion from simultaneous measurements," in *Proc of the 15th ICPhS*, 2003, pp. 431–434.

[129] A.P. Witkin, "Scale-space filtering," in *Proc. Int'l Joint Conf. on Artificial Intel.*, 1983, pp. 1019–1022.

[130] J.J. Koenderink, "The structure of images," *Biological Cybernetics*, vol. 50, pp. 363–370, 1984.

[131] P. Perona and J. Malik, "Scale space and edge detection using anisotropic diffusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, July 1990.

[132] L. Alvarez, P.L. Lions, and J.M. Morel, "Image selective smoothing and edge detection by nonlinear diffusion, II," *SIAM Journal Numer. Anal.*, vol. 29, no. 3, pp. 845–866, 1992.

[133] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[134] J. Weickert, "Multiscale texture enhancement," in *Computer Analysis of Images and Patterns*, V. Hlavác and R. Sára, Eds., vol. 970, pp. 230–237. Lecture Notes in Computer Science, Springer, Berlin, 1995.

[135] D. Tschumperlé and R. Deriche, "Vector-valued image regularization with PDE's : A common framework for different applications," in *IEEE Conf. on Comp. Vision and Pattern Recogn.*, Madison, Wisconsin (United States), June 2003.

[136] F. Guichard and F. Malgouyres, "Total variation based interpolation," *EUSIPCO III*, pp. 1741–1744, 1998.

[137] A. Belahmidi and F. Guichard, "A partial differential equation approach to image zoom," *Int. Conf. on Image Processing*, pp. 649–652, 2004.

[138] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1987.

[139] M.K. Tiede and E. Vatikiotis-Bateson, "Extracting articulator movement parameters from a videodisc-based cineradiographic database," in *Proc. of ICSLP, Yokohama*, 1994, pp. 45–48.

[140] M.-O. Berger and Y. Laprie, "Tracking Articulators in X-Ray Images with Minimal User Interaction: Example of the Tongue Extraction," in *Proceedings of IEEE International Conference on Image Processing, Lausanne, Switzerland*, September 1996, pp. 289–292.

[141] Y. Laprie and M.O. Berger, "Towards automatic extraction of tongue contours in x-ray images," in *Proceedings of International Conference on Spoken Language Processing 96*, Philadelphia (USA), October 1996, vol. 1, pp. 268–271.

[142] V. Caselles, F. Catte, T. Coll, and F. Dibos, "A geometric model for active contours," *Numer. Math.*, vol. 66, pp. 1–31, 1993.

[143] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *Int'l J. of Comp. Vision*, vol. 21, no. 1, pp. 61–79, 1997.

[144] S. Kichesammany, A. Kumar, PJ Olver, A. Tannenbaum, and A. Yezzi, "Gradient flows and geometric active contours," in *Proc. Int'l Conf. on Comp. Vision*, 1995, pp. 810–815.

[145] J. A. Sethian, *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, 2nd edition, 1999.

[146] N. Paragios and R. Deriche, "Geodesic active regions and level set methods for supervised texture segmentation," *Int'l J. of Comp. Vision*, vol. 46, no. 3, pp. 223–247, Feb. 2002.

[147] M. Leventon, E. Grimson, and O. Faugeras, "Statistical shape influence in geodesic active contours," *Proc. IEEE Conf. on Comp. Vision and Pat. Recog.*, pp. 1316–1323, 2000.

[148] N. Paragios and M. Rousson, "Shape analysis towards model-based segmentation," in *Geometric Level Set Methods in Imaging, Vision and Graphics*, 2003.

[149] D. Cremers, N. Sochen, and C. Schnorr, "Multiphase dynamic labeling for variational recognition-driven image segmentation," in *Proc. European Conf. on Comp. Vision*, 2004, pp. 74–86.

[150] G. Thimm and J. Luettin, "Extraction of articulators in x-ray image sequences," in *Proc. EUROSPEECH*, Budapest, september 1999, pp. 157–160.

[151] J. Fontecave and F. Berthommier, "Quasi-automatic extraction method of tongue movement from a large existing speech cineradiographic database ," in *Interspeech, Lisboa*, Sept. 2005, pp. 1081–1084.

[152] J. Fontecave and F. Berthommier, "Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data," in *Interspeech, Pittsburgh*, Sept. 2006, pp. 569–573.

[153] J. Serra, *Image Analysis and Mathematical Morphology*, Academic Press, NY, 1982.

[154] J. Serra, *Image Analysis and Mathematical Morphology*, vol. 2, Academic Press, NY, 1988.

[155] P. Maragos and R. W. Schafer, "Morphological systems for multidimensional signal processing," *Proc. IEEE*, vol. 78, pp. 690–710, Apr. 1990.

[156] H.J.A.M. Heijmans, *Morphological Image Operators*, Acad. Press, Boston, 1994.

[157] P. Maragos, "Morphological filtering for image enhancement and detection," in *The Image and Video Processing Handbook*, A. C. Bovik, Ed., pp. 135–156. Elsevier Acad. Press, second edition, 2005.

[158] S. Beucher and F. Meyer, "The morphological approach to segmentation: The watershed transformation," in *Mathematical Morphology in Image Processing*, E. R. Dougherty, Ed., pp. 433–481. Marcel Dekker, New York, 1993.

[159] L. Vincent and P. Soille, "Watershed in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intellig.*, vol. 13, pp. 583–598, June 1991.

[160] F. Meyer and P. Maragos, "Multiscale morphological segmentations based on watershed, flooding, and eikonal PDE," in *Lec. Notes Computer Science vol. 1682 (Proc. Scale-Space'99)*. 1999, pp. 351–362, Springer-Verlag.

[161] F. Meyer and P. Maragos, "Nonlinear scale-space representation with morphological levelings," *J. Visual Communic. and Image Representation*, vol. 11, pp. 245–265, 2000.

[162] A. Sofou, G. Evangelopoulos, and P. Maragos, "Coupled geometric and texture PDE-based segmentation," *Proc. Int'l Conf. Image Processing (ICIP-2005), Genoa, Italy*, pp. 650–653, Sept. 2005.

[163] P. Maragos, "Partial differential equations for morphological scale-spaces and eikonal applications," in *The Image and Video Processing Handbook*, A. C. Bovik, Ed., pp. 587–612. Elsevier Acad. Press, second edition, 2005.

[164] F. Guichard and J.M. Morel, *Image Analysis and P.D.E.s*, IPAM GBM Tutorials, 2001.

[165] F. Cao, *Geometric Curve Evolution and Image Processing*, Springer-Verlag, 2003.

[166] N. Paragios and R. Deriche, "Geodesic active regions: A new paradigm to deal with frame partition problems in computer vision," *Journal of Visual Communication and Image Representation*, 2002.

[167] S. Osher and J. A. Sethian, "Fronts propagating with curvature-dependent speed: Algorithms based on hamilton-jacobi formulations," *Journal of Computational Physics*, vol. 79, pp. 12–49, 1988.

[168] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[169] S. Sclaroff and J. Isidoro, "Active blobs: region-based, deformable appearance models," *Comput. Vis. Image Underst.*, vol. 89, pp. 197–225, 2003.

[170] M. J. Jones and T. Poggio, "Multidimensional morphable models: A framework for representing and matching object classes," *Int'l J. of Comp. Vision*, vol. 22, no. 2, pp. 107–131, 1998.

[171] G.D. Hager and P.N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 10, pp. 1025–1039, 1998.

[172] M. J. Black and A. D. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int'l J. of Comp. Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[173] D. DeCarlo and D. Metaxas, "Optical flow constraints on deformable models with applications to face tracking," *Int'l J. of Comp. Vision*, vol. 38, no. 2, pp. 99–127, 2000.

[174] V. T. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[175] M. Gleicher, "Projective registration with difference decomposition," in *Proc. IEEE Conf. on Comp. Vision and Pat. Recog.*, 1997, pp. 331–337.

[176] I. Matthews and S. Baker, "Active appearance models revisited," *Int'l J. of Comp. Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[177] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical recipes in C*, Camb. Univ. Press, 1992.

[178] S. Baker and I. Matthews, "Equivalence and efficiency of image alignment algorithms," in *Proc. IEEE Conf. on Comp. Vision and Pat. Recog.*, 2001, vol. 1, pp. 1090–1097.

[179] H.-Y. Schum and R. Szeliski, "Construction of panoramic image mosaics with global and local alignment," *Int'l J. of Comp. Vision*, vol. 16, no. 1, pp. 63–84, 2000.

[180] S. Baker, R. Gross, and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework - Part 3," Tech. Rep. CMU-RI-TR-03-35, Robotics Institute, Carnegie-Mellon University, 2003.

[181] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Comp.*, vol. 23, pp. 1080–1093, 2005.

[182] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *ICASSP*, 1994, pp. 669–672.

[183] R. Kaucic and A. Blake, "Accurate, real-time, unadorned lip tracking," in *ICCV*, 1998, pp. 370–375.

[184] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[185] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[186] R. Seymour, J. Ming, and D. Stewart, "A new posterior based audio-visual integration method for robust speech recognition," in *Interspeech*, 2005, pp. 1229–1232.

[187] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *ICCV*, 2005, pp. 1424–1431.

[188] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *IJCV*, vol. 29, no. 1, pp. 5–28, 1998.

[189] A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer Verlag, New York, NY, USA, 2001.

[190] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley & Sons, 2001.

[191] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 974–989, 1999.

[192] M. S. Gray, J. R. Movellan, and T. J. Sejnovski, "A comparison of image processing techniques for visual speech recognition applications," in *NIPS*, 2000, pp. 939–945.

[193] T-K. Kim, H. Kim, W. Hwang, and J. Kittler, "Independent component analysis in a local residue space for face recognition," *Pattern Recognition*, vol. 2004, no. 37, pp. 1873–1885, 2004.

[194] R. Beichel, H. Bischof, F. Leberl, and M. Sonka, "Robust active appearance models and their application to medical image analysis," *IEEE Tr. on Med. Imag.*, vol. 24, no. 9, pp. 1151–1169, 2005.

[195] T. Niikawa, M. Matsumura, T. Tachimura, and T. Wada, "Modeling of a speech production system based on MRI measurement of three-dimensional vocal tract shapes during fricative consonant phonation," in *Proceedings of the* 6^{th} *ICSLP*, 2000, vol. 2, pp. 174–177.

[196] H. Matsuzaki, N. Miki, N Nagai, T. Hirohku, and Y. Ogawa, "3D fem analysis of vocal tract model of elliptic tube with inhomogenous-wall impedance," in *Proceedings of the* 3^{rd} *ICSLP*, 1994, pp. 635–638.

[197] H. Matsuzaki and K. Motoki, "Fem analysis of 3-d vocal tract shape model with asymmetrical shape," in *Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, 2000, pp. 329–332.

[198] D. Sinder, G. Richard, H. Duncan, Q. Lin, J. Flanagan, M. Krane, S. Levinson, D. Davis, and S. Slimon, "A fluid flow approach to speech generation," in *Proceedings of the* 1^{st} *ESCA Tutorial and Research Workshop on Speech Production Modeling –* 4^{th} *Speech Production Seminar*, 1996, pp. 203–206.

[199] K. N. Stevens and S. A. House, "Development of a quantitative description of vowel articulation," *J. Acoust. Soc. Am*, vol. 27, pp. 484–493, 1955.

[200] G. Fant, *Acoustic theory of speech production*, Mouton & Co., The Hague, 1960.

[201] K. N. Stevens, "The quantal nature of speech: Evidence from articulatory-acoustic data," in *Human communication: A unified view*, E. E. David and P. B. Denes, Eds., pp. 51–66. McGraw-Hill Book Company, 1972.

[202] M. Mrayati, R. Carré, and B. Guerin, "Distinctive regions and modes : A new theory of speech production," *Speech Communication*, vol. 7, pp. 257–286, 1988.

[203] R. Carré and M. Mrayati, "Vowel-vowel trajectories and region modelling," *Journal of Phonetics*, vol. 19, pp. 433–443, 1991.

[204] C. H. Coker and O. Fujimura, "Model for specification of the vocal-tract area function," *J. Acoust. Soc. Am*, vol. 40, pp. 1271, 1966.

[205] C. H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis from ordinary english text," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 293–298, 1973.

[206] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoust. Soc. Am*, vol. 53, pp. 1070–1082, 1973.

[207] B. Lindblom and J. Sundberg, "Acoustic consequences of lip, tongue, jaw, and larynx movement," *J. Acoust. Soc. Am*, vol. 50, pp. 1166–1179, 1971.

[208] S. Parthasarathy and C. H. Coker, "On automatic estimation of articulatory parameters in a text-to-speech system," *Computer Speech and Language*, vol. 6, pp. 37–75, 1992.

[209] J. M. Heinz and K. N. Stevens, "On the relations between lateral cineradiographs, area function, and acoustic spectra of speech," in *Fifth International Congress of Acoustics, Lièges*, 1965, p. A44.

[210] J. Liljencrants, "Fourier series description of the tongue profile," Tech. Rep. QPSR 4, Speech Transmission Lab. KTH (Stockholm), 1971.

[211] R. Harshman, P. Ladefoged, and L. Goldstein, "Factor analysis of tongue shapes," *J. Acoust. Soc. Am*, vol. 62, pp. 693–707, 1977.

[212] S. Kiritani, S. Sekimoto, and H. Imagawa, "Parameter description of the tongue movements for vowels," Tech. Rep. Ann. Bull. RILP, 11, University of Tokyo, 1977.

[213] S. Maeda, "Un modèlearticulatoire de la langue avec des comosantes linéaires," in *10ièmes Journées d'Etudes sur la Parole, GALF*, 1978, pp. 152–164.

[214] A. Bothorel, P. Simon, F. Wioland, and J.-P. Zerling, "Cinéradiographie des voyelles et consonnes du français," Tech. Rep., Institut de Phonétique de Strasbourg (France), 1986.

[215] S. Maeda, "Compensatory articulation during speech : evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech production and speech modeling (NATO Advanced Study Institute series)*, W.J. Hardcastle and A. Marchal, Eds., pp. 131–149. Kluwer Academic Publishers, 1990.

[216] J. E. Overall, "Orthogonal factors and uncorrelated facor scores," *Psychological Reports*, vol. 10, pp. 651–662, 1962.

[217] T. Baer, J. C. Gore, L. C. Graco, and P. W. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *J. Acoust. Soc. Am*, vol. 90, pp. 799–828, 1991.

[218] P. Perrier, L. J. Boe, and R. Sock, "Vocal tract area function estimation from migsagittal dimensions with ct scans and a vocal tract cast," *Journal of Speech and Hearing Research*, vol. 35, pp. 53–67, 1992.

[219] S. Maeda, "Conversion of midsagittal dimensions to vocal tact area function," *J. Acoust. Soc. Am*, vol. 51, pp. 88, 1972.

[220] U. G. Goldstein, *An articulatory model for the vocal-tracts of growing children*, 1980, Ph.D. thesis MIT, Camblidge, Mass. USA.

[221] S. Bouabana and S. Maeda, "Multipulse LPC modeling of articulatory movements," *Speech Communication*, vol. 24, pp. 227–248, 1998.

[222] K. N. Stevens, "On the quantal nature of speech," *Journal of Phonetics*, vol. 17, pp. 3–45, 1989.

[223] J. L. Flanagan, *Speech analysis synthesis and perception (2nd edition)*, Springer-Verlag, New York, Heifelberg, Berlin, 1972.

[224] C. H. Shadle, *The acoustics of fricative consonants*, 1985, Ph.D. thesis MIT, Camblidge, Mass. USA.

[225] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *J. Acoust. Soc. Am*, vol. 50, no. 4, pp. 1180–1192, 1971.

[226] K. N. Stevens, "Modelling affricate consonants," *Speech Communication*, vol. 13, pp. 33–43, 1993.

[227] M. Rothenberg, *The breath-stream dynamics of simple-released-plosive production*, Number 6. Karger, Basel, 1968.

[228] C. Scally, "Speech production simulated with a functional model of the larynx and the vocal tract," *Journal of Phonetics*, vol. 14, pp. 407–414, 1986.

[229] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell System Technical Journal*, vol. 50, no. 6, pp. 1233–1268, 1972.

[230] G. Fant, "Glottal source and excitation analysis," Tech. Rep. STL-QPSR 1, KTH, Stockholm, 1979.

[231] M. Yeou and S. Maeda, "Pharyngeal and uvular consonants are approximants: An acoustic modeling study," in *The 13-th International Congress of Phonetic Sciences*, 1995, vol. 2, pp. 586–589.

[232] L. R. Rabiner and A. E. Rosenberg, "Considerations in dynamic time warping algorithm for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 6, december 1978.

[233] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc, 1975.

[234] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, pp. S35, 1975.

[235] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using chebyshev polynomials," *IEEE Trans. ASSP*, vol. 34, no. 6, pp. 1419–1426, 1986.

[236] O. Engwall, "Articulatory synthesis using corpus-based estimation of line spectrum pairs," in *Interspeech, Lisboa*, Sept. 2005, pp. 1909–1912.

[237] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Trans. IECE*, vol. J62-A, no. 4, pp. 217–223, 1979 (in Japanese).

[238] T. Gallas and X. Rodet, "Generalized functional approximation for source-filter system modelling," in *Proceedings of European Conference on Speech Technology*, Genova, Italy, September, 1991.

[239] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, August 1980.

[240] H. Hermansky, "Perceptual linear predictive (lpl) analysis of speech," *Journal of Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[241] W. Ding and N. Campbell, "Optimising unit selection with voice source and formants in the chatr speech synthesis system," in *Proc. Eurospeech*, 1997, pp. 537–540.

[242] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *Journal of the Acoustical Society of America*, vol. 108, no. 6, pp. 3036–3048, 2000.

[243] J Schoentgen and S. Ciocea, "Kinematic formant-to-area mapping," *Speech Communication*, vol. 21, no. 4, pp. 227–244, 1997.

[244] J. L. Flanagan, "Automatic extraction of formant frequencies from continuous speech," *Journal of Acoustical Society of America*, vol. 50, pp. 110–118, 1956.

[245] R.W. Schafer and L.R. Rabiner, "System for automatic formant analysis of voiced speech," *Journal of Acoustical Society of America*, vol. 47, no. 2, pp. 634–648, February 1970.

[246] S. S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 2, pp. 135–141, April 1974.

[247] L. Deng, L. Lee, H. Attias, and A. Acero, "Adaptive kalman filtering and smoothing for tracking vocal tract resonances using a continuous valued hidden dynamic model," *IEEE Transaction on Audio, Speech and Language Processing*, 2006.

[248] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. of Int. Conf. on Audio, Speech and Signal Processing, ICASSP*, 2004.

[249] R. Togneri and L. Deng, "A state-space model with neural-network prediction for recovering vocal tract resonances in fluent speech from mel-cepstral coefficients," *Speech Communication*, 2006.

[250] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 425–434, March 2006.

[251] D. Toledano, J. Villardebo, and L. Gomez, "Initialization, training and context-dependency in hmm-based formant tracking," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 14, pp. 511–523, March 2006.

[252] M. Lee, J. van Santen, B. Moebius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Transactions Using Context-Dependent Phonemic Information*, vol. 13, no. 5, pp. 741–750, September 2005.

[253] K. Mustafa and I. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 435–443, March 2006.

[254] Ramdas Kumaresan Ashwin Rao, "On decomposing speech into modulated components," *IEEE Transactions on speech and audio processing*, vol. 8, no. 3, pp. 240–254, May 2000.

[255] Yves Laprie, "A concurrent curve strategy for formant tracking," in *Proc. Int. Conf. Speech and Language Processing*, 2004.

[256] Y. Laprie and M.-O. Berger, "Cooperation of regularization and speech heuristics to control automatic formant tracking," *Speech Communication*, vol. 19, no. 4, pp. 255–270, October 1996.

[257] B. Bozkurt, B. Doval, C. D'Alessandro, and T. Dutoit, "Improved differential phase spectrum processing for formant tracking," in *Proc. of Int. Conf. on Speech, Language and Signal Processing*, 2004.

[258] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J.Acoust.Soc.Am.*, vol. 99, no. 6, pp. 3795–3806, June 1996.

[259] K. Xia and C. Espy-Wilson, "A new strategy of formant tracking based on dynamic programming," in *Proc. International Conference on Speech and Language Processing, ICSLP*, 2000.

[260] S. Krstulovic, *Speech Analysis with production constraints, Ph. D. Thesis*, Ecole Polytechnique Fédérale, Lausanne, Switzerland, 2001.

[261] H. Wakita and A. Gray, "Numerical determination of the lip impedance and vocal tract area functions," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 23, no. 6, pp. 574–580, 1975.

[262] P. Mokhtari, *An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy*, University of New South Wales, June 1998.

[263] S. L. Campbell and C. D. Meyer, *Generalized Inverses of Linear Transformations*, Dover Publications, 1991.

[264] J. Schoentgen and S. Ciocea, "Kinematic formant-to-area mapping," *Speech Communication*, vol. 21, pp. 227–244, 1997.

[265] S. Ciocea, *Semi-analytic fornant-to-area mapping, PhD Thesis*, Université Libre de Bruxelles, Brussels, Belgium, 1997.

[266] C. H. Coker, "A model of articulary dynamics and control," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 452–460, 1976.

[267] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Communication*, vol. 5, no. 2, pp. 159–170, 1986.

[268] K. Shirai and T. Kobayashi, "Estimation of articulatory motion using neural networks," *J. Phon*, vol. 19, pp. 379–385, 1991.

[269] V. N. Sorokin, "Inverse problems for fricatives," *Speech Communication*, vol. 14, pp. 249–262, 1994.

[270] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Trans. ASSP*, vol. 2, no. 1, Part. II, pp. 133–150, January 1994.

[271] K. N. Stevens and A. S. House, "Development of a quantitative description of vowel articulation," *Journal of the Acoustical Society of America*, vol. 27, pp. 484–493, 1955.

[272] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Trans. ASSP*, vol. 36, no. 12, pp. 1812–1818, December 1988.

[273] V.N. Sorokin and A.V. Trushkin, "Articulatory-to-acoustic mapping for inverse problem," *Speech Communication*, vol. 19, pp. 105–118, 1996.

[274] Schroeter J., Meyer P., and Parthasarathy S., "Evaluation of improved articulatory codebooks and codebook access distance measures," in *Proc. ICASSP*, Albuquerque, NM, USA, April 1990, pp. 393–396.

[275] S. Ouni and Y. Laprie, "Utilisation d'un dictionnaire hypercubique pour l'inversion acoustico-articulatoire," in *Actes des Journées d'Étude sur la parole, Aussois*, June 2000, pp. 409–412.

[276] A. Soquet, M. Saerens, and P. Jospa, "Acoustic-articulatory inversion based on a neural controller of a vocal tract model," in *Proceedings of the ESCA workshop on speech synthesis*, Autrans, France, sep 1990.

[277] M.G. Rahim, W.B. Kleijn, J. Schroeter, and C.C. Goodyear, "Acoustic to articulatory parameter mapping using an assembly of neural networks," in *Proc. ICASSP*, Toronto, may 1991, pp. 485–488.

[278] K. Richmond, "A trajectory mixture density network for the acoustic-articulatory inversion mapping," in *Proc. of Interspeech*, Pittsburgh, USA, september 2006.

[279] T. Toda, A. W. Black, and K. Tokuda, "Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model," in *Proc. ICSLP*, Jeiju, Korea, Oct. 2004.

[280] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, Mars 2004.

[281] S. Maeda, "Un modèle articulatoire de la langue avec des composantes linéaires," in *Actes 10èmes Journées d'Etude sur la Parole*, Grenoble, Mai 1979, pp. 152–162.

[282] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, pp. 1070–1082, 1973.

[283] J. Hogden, P. Rubin, and E. Saltzman, "An unsupervised method for learning to track tongue position from an acoustic signal," *Les cahiers de l'I.C.P., Bulletin de la communication parlée*, vol. 3, pp. 101–116, 1995.

[284] Potard B. and Laprie Y., "Using phonetic constraints in acoustic-to-aticulatory inversion," in *Interspeech, Lisboa*, Sept. 2005, pp. 3217–3220.

[285] S. Dusan and L. Deng, "Recovering vocal tract shapes from MFCC parameters," in *Proc. ICSLP*, Sydney (Australia), December 1998, vol. 2, pp. 3087–3090.

[286] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds., pp. 231–267. Dekker, New York, 1992.

[287] S. Dusan, "Methods for Integrating Phonetic and Phonological Knowledge in Speech Inversion," in *Proceedings of the International Conference on Speech, Signal and Image Processing*, Malta, September 2001.

[288] Y. Laprie and B. Mathieu, "A variational approach for estimating vocal tract shapes from the speech signal," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, May 1998, vol. 2, pp. 929–932.

[289] O. Engwall, "Evaluation of speech inversion using an articulatory classifier," in *Proceedings of the 7^{th} International Seminar on Speech Production*, 2006, pp. 469–476.

[290] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175–185, March 2004.

[291] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[292] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 133–150, 1992.

[293] S. Roweis, "Constrained hidden markov models," *In Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, vol. 12, 2000.

[294] G. Bailly and P. Badin, "Seeing the tongue from outside," in *Proc of the 6^{th} ICSLP*, Denver, Sept. 2002, pp. 1913–1916.

[295] O. Engwall and J. Beskow, "Resynthesis of 3D tongue movements from facial data," in *Proc of Eurospeech*, 2003, pp. 2261–2264.

[296] D.W. Massaro, *Speech reading by Ear and Eye*, Erlaub, Hillsdale, 1987.