MORGAN & CLAYPOOL PUBLISHERS

Dynamic Speech Models

Theory, Algorithms, and Applications

Li Deng

SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING

Dynamic Speech Models

Theory, Algorithms, and Applications

Li Deng
Microsoft Research
Redmond, Washington, USA

SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING #2

MORGAN & CLAYPOOL PUBLISHERS

Dynamic Speech Models, Theory, Algorithms, and Applications

Li Deng

www.morganclaypool.com

1598290649 paper Deng
1598290657 ebook Deng

DOI: 10.2200/S00028ED1V01Y200605SAP002

A Publication in the Morgan & Claypool Publishers' series
SYNTHESIS LECTURES ON SPEECH AND AUDIO PROCESSING
Lecture #2
Series editor B. H. Juang

First Edition

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

Dynamic Speech Models

Theory, Algorithms, and Applications

ABSTRACT

Speech dynamics refer to the temporal characteristics in all stages of the human speech communication process. This speech "chain" starts with the formation of a linguistic message in a speaker's brain and ends with the arrival of the message in a listener's brain. Given the intricacy of the dynamic speech process and its fundamental importance in human communication, this monograph is intended to provide a comprehensive material on mathematical models of speech dynamics and to address the following issues: How do we make sense of the complex speech process in terms of its functional role of speech communication? How do we quantify the special role of speech timing? How do the dynamics relate to the variability of speech that has often been said to seriously hamper automatic speech recognition? How do we put the dynamic process of speech into a quantitative form to enable detailed analyses? And finally, how can we incorporate the knowledge of speech dynamics into computerized speech analysis and recognition algorithms? The answers to all these questions require building and applying computational models for the dynamic speech process.

What are the compelling reasons for carrying out dynamic speech modeling? We provide the answer in two related aspects. First, scientific inquiry into the human speech code has been relentlessly pursued for several decades. As an essential carrier of human intelligence and knowledge, speech is the most natural form of human communication. Embedded in the speech code are linguistic (as well as para-linguistic) messages, which are conveyed through four levels of the speech chain. Underlying the robust encoding and transmission of the linguistic messages are the speech dynamics at all the four levels. Mathematical modeling of speech dynamics provides an effective tool in the scientific methods of studying the speech chain. Such scientific studies help understand why humans speak as they do and how humans exploit redundancy and variability by way of multitiered dynamic processes to enhance the efficiency and effectiveness of human speech communication. Second, advancement of human language technology, especially that in automatic recognition of natural-style human speech is also expected to benefit from comprehensive computational modeling of speech dynamics. The limitations of current speech recognition technology are serious and are well known. A commonly acknowledged and frequently discussed weakness of the statistical model underlying current speech recognition technology is the lack of adequate dynamic modeling schemes to provide correlation structure across the temporal speech observation sequence. Unfortunately, due to a variety of reasons, the majority of current research activities in this area favor only incremental modifications and improvements to the existing HMM-based state-of-the-art. For example, while the dynamic and correlation modeling is known to be an important topic, most of the systems nevertheless employ only an ultra-weak form of speech dynamics; e.g., differential or delta parameters. Strong-form dynamic speech modeling, which is the focus of this monograph, may serve as an ultimate solution to this problem.

After the introduction chapter, the main body of this monograph consists of four chapters. They cover various aspects of theory, algorithms, and applications of dynamic speech models, and provide a comprehensive survey of the research work in this area spanning over past 20 years. This monograph is intended as advanced materials of speech and signal processing for graudate-level teaching, for professionals and engineering practioners, as well as for seasoned researchers and engineers specialized in speech processing.

KEYWORDS

Articulatory trajectories, Automatic speech recognition, Coarticulation, Discretizing hidden dynamics, Dynamic Bayesian network, Formant tracking, Generative modeling, Speech acoustics, Speech dynamics, Vocal tract resonance

Acknowledgments

This book would not have been possible without the help and support from friends, family, colleagues, and students. Some of the material in this book is the result of collaborations with my former students and current colleagues. Special thanks go to Jeff Ma, Leo Lee, Dong Yu, Alex Acero, Jian-Lai Zhou, and Frank Seide.

The most important acknowledgments go to my family. I also thank Microsoft Research for providing the environment in which the research described in this book is made possible. Finally, I thank Prof. Fred Juang and Joel Claypool for not only the initiation but also the encouragement and help throughout the course of writting this book.

Contents

CHAPTER 1

Introduction

1.1 WHAT ARE SPEECH DYNAMICS?

In a broad sense, speech dynamics are time-varying or temporal characteristics in all stages of the human speech communication process. This process, sometimes referred to as speech chain [1], starts with the formation of a linguistic message in the speaker's brain and ends with the arrival of the message in the listener's brain. In parallel with this direct information transfer, there is also a feedback link from the acoustic signal of speech to the speaker's ear and brain. In the conversational mode of speech communication, the style of the speaker's speech can be further influenced by an assessment of the extent to which the linguistic message is successfully transferred to or understood by the listener. This type of feedbacks makes the speech chain a closed-loop process.

The complexity of the speech communication process outlined above makes it desirable to divide the entire process into modular stages or levels for scientific studies. A common division of the direct information transfer stages of the speech process, which this book is mainly concerned with, is as follows:

- *Linguistic level*: At this highest level of speech communication, the speaker forms the linguistic concept or message to be conveyed to the listener. That is, the speaker decides to say something linguistically meaningful. This process takes place in the language center(s) of speaker's brain. The basic form of the linguistic message is words, which are organized into sentences according to syntactic constraints. Words are in turn composed of syllables constructed from phonemes or segments, which are further composed of phonological features. At this linguistic level, language is represented in a discrete or symbolic form.
- *Physiological level*: Motor program and articulatory muscle movement are involved at this level of speech generation. The speech motor program takes the instructions, specified by the segments and features formed at the linguistic level, on how the speech sounds are to be produced by the articulatory muscle (i.e., articulators) movement over time. Physiologically, the motor program executes itself by issuing time-varying commands imparting continuous motion to the articulators including the lips, tongue,

larynx, jaw, and velum, etc. This process involves coordination among various articulators with different limitations in the movement speed, and it also involves constant corrective feedback. The central scientific issue at this level is how the transformation is accomplished from the discrete linguistic representation to the continuous articulators' movement or dynamics. This is sometimes referred to as the problem of interface between phonology and phonetics.

- *Acoustic level*: As a result of the articulators' movements, acoustic air stream emerges from the lungs, and passes through the vocal cords where a phonation type is developed. The time-varying sound sources created in this way are then filtered by the time-varying acoustic cavities shaped by the moving articulators in the vocal tract. The dynamics of this filter can be mathematically represented and approximated by the changing vocal tract area function over time for many practical purposes. The speech information at the acoustic level is in the form of dynamic sound pattern after this filtering process. The sound wave radiated from the lips (and in some cases from the nose and through the tissues of the face) is the most accessible element of the multiple-level speech process for practical applications. For example, this speech sound wave may be easily picked by a microphone and be converted to analog or digital electronic form for storage or transmission. The electronic form of speech sounds makes it possible to transport them thousands of miles away without loss of fidelity. And computerized speech recognizers gain access to speech data also primarily in the electronic form of the original acoustic sound wave.
- *Auditory and perceptual level*: During human speech communication, the speech sound generated at the acoustic level above impinges upon the eardrums of a listener, where it is first converted to mechanical motion via the ossicles of the middle ear, then to fluid pressure waves in the medium bathing the basilar membrane of the inner ear invoking traveling waves. This finally excites hair cells' electrical, mechanical, and biochemical activities, causing firings in some 30,000 human auditory nerve fibers. These various stages of the processing carry out some nonlinear form of frequency analysis, with the analysis results in the form of dynamic spatial–temporal neural response patterns. The dynamic spatial–temporal neural responses are then sent to higher processing centers in the brain, including the brainstem centers, the thalamus, and the primary auditory cortex. The speech representation in the primary auditory cortex (with a high degree of plasticity) appears to be in the form of multiscale and jointly spectro-temporally modulated patterns. For the listener to extract the linguistic content of speech, a process that we call speech perception or *decoding*, it is necessary to identify the segments and features that underlie the sound pattern based on the speech representation in the

primary auditory cortex. The decoding process may be aided by some type of analysis-by-synthesis strategies that make use of general knowledge of the dynamic processes at the physiological and acoustic levels of the speech chain as the "encoder" device for the intended linguistic message.

At all the four levels of the speech communication process above, dynamics play a central role in shaping the linguistic information transfer. At the linguistic level, the dynamics are discrete and symbolic, as is the phonological representation. That is, the discrete phonological symbols (segments or features) change their identities at various points of time in a speech utterance, and no quantitative (numeric) degree of change and precise timing are observed. This can be considered as a weak form of dynamics. In contrast, the articulatory dynamics at the physiological level, and the consequent dynamics at the acoustic level, are of a strong form in that the numerically quantifiable temporal characteristics of the articulator movements and of the acoustic parameters are essential for the trade-off between overcoming the physiological limitations for setting the articulators' movement speed and efficient encoding of the phonological symbols. At the auditory level, the importance of timing in the auditory nerve's firing patterns and in the cortical responses in coding speech has been well known. The dynamic patterns in the aggregate auditory neural responses to speech sounds in many ways reflect the dynamic patterns in the speech signal, e.g., time-varying spectral prominences in the speech signal. Further, numerous types of auditory neurons are equipped with special mechanisms (e.g., adaptation and onset-response properties) to enhance the dynamics and information contrast in the acoustic signal. These properties are especially useful for detecting certain special speech events and for identifying temporal "landmarks" as a prerequisite for estimating the phonological features relevant to consonants [2, 3].

Often, we use our intuition to appreciate speech dynamics—as we speak, we sense the motions of speech articulators and the sounds generated from these motions as continuous flow. When we call this continuous flow of speech organs and sounds as speech dynamics, then we use them in a narrow sense, ignoring their linguistic and perceptual aspects.

As is often said, timing is of essence in speech. The dynamic patterns associated with articulation, vocal tract shaping, sound acoustics, and auditory response have the key property that the timing axis in these patterns is adaptively plastic. That is, the timing plasticity is flexible but not arbitrary. Compression of time in certain portions of speech has a significant effect in speech perception, but not so for other portions of the speech. Some compression of time, together with the manipulation of the local or global dynamic pattern, can change perception of the style of speaking but not the phonetic content. Other types of manipulation, on the other hand, may cause very different effects. In speech perception, certain speech events, such as labial stop bursts, flash extremely quickly over as short as 1–3 ms while providing significant cues for the listener

to identify the relevant phonological features. In contrast, for other phonological features, even dropping a much longer chunk of the speech sound would not affect their identification. All these point to the very special status of time in speech dynamics. The time in speech seems to be quite different from the linear flow of time as we normally experience it in our living world.

Within the speech recognition community, researchers often refer to speech dynamics as differential or regression parameters derived from the acoustic vector sequence (called delta, delta–delta, or "dynamic" features) [4, 5]. From the perspective of the four-level speech chain outlined above, such parameters can at best be considered as an ultra-weak form of speech dynamics. We call them ultra-weak not only because they are confined to the acoustic domain (which is only one of the several stages in the complete speech chain), but also because temporal differentiation can be regarded hardly as a full characterization in the actual dynamics even within the acoustic domain. As illustrated in [2, 6, 7], the acoustic dynamics of speech exhibited in spectrograms have the intricate, linguistically correlated patterns far beyond what the simplistic differentiation or regression can characterize. Interestingly, there have been numerous publications on how the use of the differential parameters is problematic and inconsistent within the traditional pattern recognition frameworks and how one can empirically remedy the inconsistency (e.g., [8]). The approach that we will describe in this book gives the subject of dynamic speech modeling a much more comprehensive and rigorous treatment from both scientific and technological perspectives.

1.2 WHAT ARE MODELS OF SPEECH DYNAMICS?

As discussed above, the speech chain is a highly dynamic process, relying on the coordination of linguistic, articulatory, acoustic, and perceptual mechanisms that are individually dynamic as well. How do we make sense of this complex process in terms of its functional role of speech communication? How do we quantify the special role of speech timing? How do the dynamics relate to the variability of speech that has often been said to seriously hamper automatic speech recognition? How do we put the dynamic process of speech into a quantitative form to enable detailed analyses? How can we incorporate the knowledge of speech dynamics into computerized speech analysis and recognition algorithms? The answers to all these questions require building and applying computational models for the dynamic speech process.

A computational model is a form of mathematical abstraction of the realistic physical process. It is frequently established with necessary simplification and approximation aimed at mathematical or computational tractability. The tractability is crucial in making the mathematical abstraction amenable to computer or algorithmic implementation for practical engineering applications. Applying this principle, we define *models of speech dynamics* in the context of this book as the mathematical characterization and abstraction of the physical speech dynamics. These characterization and abstraction are capable of capturing the essence of time-varying

aspects in the speech chain and are sufficiently simplified to facilitate algorithm development and engineering system implementation for speech processing applications. It is highly desirable that the models be developed in statistical terms, so that advanced algorithms can be developed to automatically and optimally determine any parameters in the models from a representative set of training data. Further, it is important that the probability for each speech utterance be efficiently computed under any hypothesized word-sequence transcript to make the speech decoding algorithm development feasible.

Motivated by the multiple-stage view of the dynamic speech process outlined in the preceding section, detailed computational models, especially those for the multiple generative stages, can be constructed from the distinctive feature-based linguistic units to acoustic and auditory parameters of speech. These stages include the following:

- A discrete feature-organization process that is closely related to speech gesture overlapping and represents partial or full phone deletion and modifications occurring pervasively in casual speech;
- a segmental target process that directs the model-articulators up-and-down and front-and-back movements in a continuous fashion;
- the target-guided dynamics of model-articulators movements that flow smoothly from one phonological unit to the next; and
- the static nonlinear transformation from the model-articulators to the measured speech acoustics and the related auditory speech representations.

The main advantage of modeling such detailed multiple-stage structure in the dynamic human speech process is that a highly compact set of parameters can then be used to capture phonetic context and speaking rate/style variations in a unified framework. Using this framework, many important subjects in speech science (such as acoustic/auditory correlates of distinctive features, articulatory targets/dynamics, acoustic invariance, and phonetic reduction) and those in speech technology (such as modeling pronunciation variation, long-span context-dependence representation, and speaking rate/style modeling for recognizer design) that were previously studied separately by different communities of researchers can now be investigated in a unified fashion.

Many aspects of the above multitiered dynamic speech model class, together with its scientific background, have been discussed in [9]. In particular, the feature organization/overlapping process, as is central to a version of computational phonology, has been presented in some detail under the heading of "computational phonology." Also, some aspects of auditory speech representation, limited mainly to the peripheral auditory system's functionalities, have been elaborated in [9] under the heading of "auditory speech processing." This book will treat these

topics only lightly, especially considering that both computational phonology and high-level auditory processing of speech are still active ongoing research areas. Instead, this book will concentrate on the following:

- The target-based dynamic modeling that interfaces between phonology and articulation-based phonetics;
- the switching dynamic system modeling that represents the continuous, target-directed movement in the "hidden" articulators and in the vocal tract resonances being closely related to the articulatory structure; and
- the relationship between the "hidden" articulatory or vocal tract resonance parameters to the measurable acoustic parameters, enabling the hidden speech dynamics to be mapped stochastically to the acoustic dynamics that are directly accessible to any machine processor.

In this book, these three major components of dynamic speech modeling will be treated in a much greater depth than in [9], especially in model implementation and in algorithm development. In addition, this book will include comprehensive reviews of new research work since the publication of [9] in 2003.

1.3 WHY MODELING SPEECH DYNAMICS?

What are the compelling reasons for carrying out dynamic speech modeling? We provide the answer in two related aspects. First, scientific inquiry into the human speech code has been relentlessly pursued for several decades. As an essential carrier of human intelligence and knowledge, speech is the most natural form of human communication. Embedded in the speech code are linguistic (and para-linguistic) messages, which are conveyed through the four levels of the speech chain outlined earlier. Underlying the robust encoding and transmission of the linguistic messages are the speech dynamics at all the four levels (in either a strong form or a weak form). Mathematical modeling of the speech dynamics provides one effective tool in the scientific methods of studying the speech chain—observing phenomena, formulating hypotheses, testing the hypotheses, predicting new phenomena, and forming new theories. Such scientific studies help understand why humans speak as they do and how humans exploit redundancy and variability by way of multitiered dynamic processes to enhance the efficiency and effectiveness of human speech communication.

Second, advancement of human language technology, especially that in automatic recognition of natural-style human speech (e.g., spontaneous and conversational speech), is also expected to benefit from comprehensive computational modeling of speech dynamics. Automatic speech recognition is a key enabling technology in our modern information society. It serves human–computer interaction in the most natural and universal way, and it also aids the

enhancement of human–human interaction in numerous ways. However, the limitations of current speech recognition technology are serious and well known (e.g., [10–13]). A commonly acknowledged and frequently discussed weakness of the statistical model (hidden Markov model or HMM) underlying current speech recognition technology is the lack of adequate dynamic modeling schemes to provide correlation structure across the temporal speech observation sequence [9, 13, 14]. Unfortunately, due to a variety of reasons, the majority of current research activities in this area favor only incremental modifications and improvements to the existing HMM-based state-of-the-art. For example, while the dynamic and correlation modeling is known to be an important topic, most of the systems nevertheless employ only the ultra-weak form of speech dynamics, i.e., differential or delta parameters. A strong form of dynamic speech modeling presented in this book appears to be an ultimate solution to the problem.

It has been broadly hypothesized that new computational paradigms beyond the conventional HMM as a generative framework are needed to reach the goal of all-purpose recognition technology for unconstrained natural-style speech, and that statistical methods capitalizing on essential properties of speech structure are beneficial in establishing such paradigms. Over the past decade or so, there has been a popular discriminant-function-based and conditional modeling approach to speech recognition, making use of HMMs (as a discriminant function instead of as a generative model) or otherwise [13, 15–19]. This approach has been grounded on the assumption that we do not have adequate knowledge about the realistic speech process, as exemplified by the following quote from [17]: "The reason of taking a discriminant function based approach to classifier design is due mainly to the fact that we lack complete knowledge of the form of the data distribution and training data are inadequate." The special difficulty of acquiring such distributional speech knowledge lies in the sequential nature of the data with a variable and high dimensionality. This is essentially the problem of dynamics in the speech data. As we gradually fill in such knowledge while pursing research in dynamic speech modeling, we will be able to bridge the gap between the discriminative paradigm and the generative modeling one, but with a much higher performance level than the systems at present. This dynamic speech modeling approach can enable us to "put speech science back into speech recognition" instead of treating speech recognition as a generic, loosely constrained pattern recognition problem. In this way, we are able to develop models "that really model speech," and such models can be expected to provide an opportunity to lay a foundation of the next-generation speech recognition technology.

1.4 OUTLINE OF THE BOOK

After the introduction chapter, the main body of this book consists of four chapters. They cover theory, algorithms, and applications of dynamic speech models and survey in a comprehensive manner the research work in this area spanning over past 20 years or so. In Chapter 2, a general framework for modeling and for computation is presented. It provides the design philosophy for dynamic speech models and outlines five major model components, including phonological