Analysis and Synthesis of Degree of Articulation

Johan Wouters
M.S. Electrical Engineering, Katholieke Universiteit Leuven (KUL), Belgium, 1996

A dissertation submitted to the faculty of the
Oregon Graduate Institute of Science and Technology
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Computer Science and Engineering

June 2001

The dissertation "Analysis and Synthesis of Degree of Articulation" by Johan Wouters has been examined and approved by the following Examination Committee:

For the late Michael W. Macon:
Dr. Jan P. H. van Santen
Professor
Thesis Research Adviser

Dr. Hynek Hermansky
Professor

Dr. James Hook
Associate Professor and Department Head

Dr. Mari Ostendorf
Professor
University of Washington

Dedication

I dedicate this thesis to my research adviser Mike Macon. Mike was a great source of inspiration and support for me during the four years that we worked together. I feel very grateful for having been his student. Mike certainly was the kindest, most resourceful and encouraging adviser one could wish for. I hope to be a little bit like him, as a researcher, as a colleague, and as a friend, so that his memory can continue to be an inspiration for me and for the people around me.

Acknowledgements

I would like to thank everyone at CSLU, now and in the past, for making this group such a stimulating place to study and to do research: Ron, for his vision and generosity; Jan, for his knowledge and support; Hynek, for setting high standards; Vincent, Alex, and Andrew, for many insightful discussions about speech synthesis; Paul, for dissecting my drafts and writing a great phonetic aligner; Rudolph, Tim, and Jacques, for their friendship and help with many questions; Charlene, for always offering a helping hand; all the native or fluent speakers of English, for participating in my perceptual tests.

I thank John and Frederik, for being such fun and tolerant roommates. I thank all my good friends in Belgium, especially Stefan, Kris, and Tom, for leading interesting lives, and for letting me be part of it. I thank Montse, for sharing with me the moments of excitement as well as frustration, which let me believe I would graduate either the next week or the next decade. I thank my parents, for always encouraging me to learn new things, for sending many emails in the past four years, and for patiently hoping that I will move a few time zones closer when I graduate.

Contents

List of Tables

List of Figures

Abstract

Analysis and Synthesis of Degree of Articulation

Johan Wouters , M.S.

Ph.D., Oregon Graduate Institute of Science and Technology
June 2001

Thesis Advisers: Michael W. Macon, Jan van Santen

Human speakers indicate the relative importance of syllables or words in an utterance through variations in intonation and speaking rate, as well as variations in the degree of articulation and vocal effort. To generate natural-sounding computer speech, corresponding acoustic variables, such as the fundamental frequency, phoneme durations, formant frequencies, and spectral balance, must be accurately controlled. However, in most speech synthesis systems, parameters related to the degree of articulation are not explicitly controlled. The resulting speech often sounds over-articulated, and requires a high listening effort as acoustic cues related to the semantic and linguistic structure of the message are missing.

The degree of articulation can be defined as the phonetic quality of a phoneme, i.e., the extent to which the target sound associated with a phoneme is realized. Moon and Lindblom have proposed a contextual model of articulation, in which the degree of articulation of vowels is described by three factors: (1) the phonetic context, (2) the phoneme duration, and (3) the spectral rate-of-change. In this thesis, we investigate whether this model can explain variations of the degree of articulation in fluent speech, and how the model may be integrated in a concatenative speech synthesis system. We focus on the spectral rate-of-change of phonetic transitions, which reflects the articulation effort used by a speaker.

We design and analyze a balanced database to study the interaction between the spectral rate-of-change and prosodic factors such as stress, accent, word position, and speaking style. A numerical model is proposed to predict the spectral rate-of-change from prosodic factors. Then, we investigate how the contextual model of articulation can be integrated in a concatenative speech synthesis system, by modifying the spectral rate-of-change of acoustic units according to the prosodic context. Spectral modification is realized using a sinusoidal + all-pole representation of the acoustic units, which avoids shortcomings of modification methods based on inverse filtering. The results show that vowel reduction and coarticulation between sonorants can be produced at synthesis time, reducing the amount of speech units needed in the acoustic inventory. Concatenation mismatch between acoustic units is also successfully diminished.

Chapter 1

Introduction

1.1 Motivation

Speech synthesis is a technology that enables computers to talk. It provides access to computer-based information in situations where people cannot read, for example to access information over the telephone, in eyes-busy applications such as driving a car, or to interact with display-less devices such as a cellphone or a microwave oven. The advantages of speech synthesis over pre-recorded voice prompts are the possibility to update information without needing new recordings, and reduced storage requirements if many prompts are needed, for example to say the time of the day or to read telephone numbers. Speech synthesis also allows people who can type but not speak to communicate verbally, and it can serve as a speech tutor for students of a foreign language or for persons with a spoken language impairment, such as the profoundly deaf (Cole *et al.*, 1998).

While existing synthesis techniques produce speech that is intelligible, few people would claim that listening to computer speech is enjoyable, or would be motivated to purchase a product from a computerized voice. Therefore, in recent years, research in speech synthesis has focused on producing speech that sounds more "natural" or "human-like". The expectation is that such speech will also sound more pleasant and more engaging. Interestingly, though, not all human voices sound equally pleasant or engaging, either. When using data-driven methods to improve the naturalness of synthesized speech, speakers or "voice talents" must therefore be carefully chosen to have the desired voice quality or speaking style (Syrdal *et al.*, 1998b).

One reason why computer speech does not sound like any human speaker, is that it is difficult to achieve appropriate *prosody.* At a phonological level, prosody refers to the relative importance or *prominence* of words, syllables, and phonemes in an utterance, which is determined by syntactic, semantic, and pragmatic information (Nooteboom, 1997; Portele and Heuft, 1997). Other functions of prosody are to indicate coherent groups of speech units and discourse turns (Nooteboom, 1997; Botinis *et al.*, 2001). Predicting prosody using computer algorithms can be difficult, as it sometimes requires that the computer "knows what it is talking about". For example, in a discussion about *Mr. Jones*, this person's name should be emphasized the first time it is mentioned, but not always in later references. When *Mrs. Jones* is also introduced, the title *Mrs.* should be emphasized contrastively. Few algorithms for prosody prediction incorporate the pragmatic and semantic knowledge required to model such interactions.

At a speech production level, prosody is defined by aspects of intonation, timing, vocal effort, and articulation style. This translates at an acoustic level into variations of the fundamental frequency (pitch), phoneme durations, spectral balance, signal intensity, formant frequencies, and other factors, which can be referred to jointly as the *acoustic-prosodic* characteristics of speech. These characteristics must be accurately controlled during speech synthesis, as they convey the prosodic structure of the message (Traunmüller, 1994).

For several years, research in prosodic modeling has focused almost exclusively on the prediction of pitch contours and phoneme durations for text-to-speech synthesis (e.g., Pierrehumbert, 1981; Ross, 1994; Hirose *et al.*, 1984; Jilka *et al.*, 1999; van Santen, 1992). The predicted pitch and durations can be imposed on acoustic units via independent time and pitch-scale modifications (Moulines and Charpentier, 1990; Quatieri and McAulay, 1992). However, even when natural pitch targets and phoneme durations are available, for example from "prosody transplantation" of a recorded sentence, concatenated utterances can still sound unnatural. Therefore, additional acoustic-prosodic characteristics need to be controlled. This is demonstrated by recent approaches in concatenative synthesis where acoustic units spoken in different prosodic contexts are collected, so that a larger variety of acoustic-prosodic variations can be displayed during synthesis (Syrdal *et al.*, 2000; Bellegarda *et al.*, 2001). However, these approaches requires ever larger databases, while not improving our understanding of the effects of prosodic context on the acoustic units.

In this thesis, acoustic variations related to the *degree of articulation* of sonorant phonemes are investigated. The degree of articulation refers to the phonetic quality of a phoneme, i.e., the extent to which the target sound associated with a phoneme is realized (Ladefoged, 1975, p. 80). Human speakers emphasize parts of a spoken message by articulating them more clearly, and de-emphasize other parts by reducing their *articulation effort.* In speech synthesis systems, however, phonemes often reach invariant phonetic targets, giving the impression that an equal amount of effort is used to realize each phoneme. The resulting utterances can sound over-articulated or "choppy", and require a high listening effort (Silverman and Morgan, 1990; van Santen, 1997; Delogu *et al.*, 1998). The aim of this thesis is to understand how the degree of articulation changes in natural speech, and to develop methods to control the degree of articulation during synthesis, so that more natural and intelligible speech can be produced.

Lindblom (1963) has pointed out that vowel reduction is a characteristic of languages with heavy stress, such as English and Swedish. Possibly, the degree of articulation plays a different role in languages such as French or Japanese, which have less stress distinctions (Botinis *et al.*, 2001). The focus of this thesis, however, is on the production and the perception of the degree of articulation in American English.

1.2 Overview of Speech Synthesis Approaches

1.2.1 From Mechanical Systems to Formant Synthesis

Early synthesis approaches included mechanical systems with bellows and resonance chambers, which could produce recognizeable speech when manipulated by a practiced operator. A well-known example is the machine of baron von Kempelen, which dates from 1778. His "voice organ" consisted of a resonance chamber, the shape of which was modulated with one hand to produce various vowels. Consonants were produced with the other hand, by controlling the flow of air from the resonance chamber. An interesting review of historical speech synthesis systems is presented by Scha (1992).

A more modern synthesis approach is to create software models of the geometry and the acoustic properties of the human vocal tract, and to compute the propagation of acoustic waves through these models. Among the challenges to be addressed in this research are the accurate description of the static and dynamic characteristics of the human articulators, i.e., the vocal chords, tongue, velum, jaws, and lips (e.g., Gabioud, 1994; Perrier and Ostry, 1994; Blackburn and Young, 2000). Computational requirements also complicate the integration of such articulatory systems in practical applications.

In the 1980's, Dennis Klatt and colleagues developed the first full Text-To-Speech (TTS) synthesizer. The system converted English text into intelligible speech, by exciting electrical resonators with an impulse train or with a noise-like signal (Klatt, 1987; Allen *et al.*, 1987). The resonators describe the resonant frequencies of the vocal tract, or formants, hence the technique is called "formant synthesis". Klatt's work culminated in the commercially available DECTALK system, which is used as an assistive technology by speaking impaired people even today.

1.2.2 Concatenative Speech Synthesis

Diphone Lookup

With the dramatic improvements in computer speed and storage, today the most popular synthesis approach is to concatenate segments of recorded speech, spoken by a single speaker. Many concatenative systems use *diphone* units, which span the transition between any two phonemes in a target language. For example, in American English, about 43 phonemes can be identified, requiring a database of $43 \times 43 = 1849$ diphones (Sproat R., editor, 1998).

The diphone units are usually articulated very clearly during the recording of the database, so that invariant phonetic targets are reached at the diphone end points. This phonetic invariance allows diphones to be concatenated at the center of a phoneme, without introducing any disturbing "glitches" or *concatenation mismatches.* However, the requirement of phonetic invariance also causes the concatenated speech to sound artificially over-articulated. The signal consists of a succession of clearly articulated phonemes, which does not demonstrate any of the phonetic reduction processes observed in human speech.

Diphone units are typically recorded as part of nonsense words or in a prosodically neutral context. The speaker is requested to adopt a monotone voice quality, to increase the consistency of the acoustic units. As a result, diphones often have long durations and a fairly constant fundamental frequency. To generate speech that sounds more lively and corresponds to a target prosody, time-scale and pitch-scale modifications are performed, imposing phoneme durations and intonation patterns predicted by the TTS engine.

Independent modification of time and pitch can be achieved using methods based on time-domain pitch-synchronous overlap-add (TD-PSOLA) (Moulines and Charpentier, 1990). For large time or pitch-scale modifications (i.e., more than a scaling factor of 2), better results have been reported based on sinusoidal analysis and synthesis of speech (McAulay and Quatieri, 1986; Quatieri and McAulay, 1992; Syrdal *et al.*, 1998a; Stylianou, 2001). However, researchers have found that time and pitch-scale modifications do not change the prosody of the acoustic units convincingly (Campbell and Black, 1996), while sometimes introducing artifacts such as buzziness or a "mechanical" voice quality (Syrdal *et al.*, 1998a).

Unit Selection

A recent trend in concatenative synthesis is to collect large databases of phonetically and prosodically varied speech. At synthesis time, speech is produced by selecting units from the database that best represent the target utterance. Usually, a dynamic search algorithm is implemented to trade off acoustic costs related to (1) the transitions between concatenated units, and (2) closeness to the phonetic and prosodic target (Hunt and Black, 1996). Time and pitch-scale modifications are either avoided entirely, or are applied selectively (Donovan, 1998).

Table 1.1 illustrates the evolution of the database size for some recent concatenative synthesis approaches. The database size is driven by the designers' wish to obtain the greatest possible variety of phonetically different units, in the greatest number of prosodically different contexts. Examples of prosodic factors taken into account are: syllabic stress, word emphasis, location of phrase boundaries, and word class (Drullman and Collier, 1991; Breen and Jackson, 1998b; Syrdal *et al.*, 2000; Bellegarda *et al.*, 2001). The

Table 1.1: Increases in database size for concatenative synthesis systems.

	text material		*time*
Nakajima and Hamada (1988)	432	words	5 min
Takeda *et al.* (1992)	5,000	words	
Nakajima (1994)	1,504	words	45 min
Iwahashi and Sagisaka (1995)	5,456	words	
Tseng (1995)	1,455	words	
	+ 599	sentences	
Hunt and Black (1996)	40,000	phonemes	50 min
Donovan (1998)			45 min
Hon *et al.* (1998)	32,000	phonemes	
Balestri *et al.* (1999)	60,000	phonemes	
Syrdal *et al.* (2000)			90 min
Takano *et al.* (2001)	60,000	multi-phone units	
Bellegarda *et al.* (2001)	4×28,000	words	
	+ 3×50,797	phonemes	

combinations between these factors result in a large variety of prosodic contexts to be covered in the database.

As can be concluded from Table 1.1, no agreement has yet been reached on the optimal database size. While one easily identifies additional prosodic contexts that should be included in the database to improve synthesis quality, little progress has been made to identify prosodic contexts or phonetic sequences that should *not* be included, or for which the acoustic-prosodic characteristics overlap. It has been argued that for synthesis of unrestricted text an unrealistically large number of acoustic units must be collected, as the interactions between prosodic factors introduce a large number of contexts to be covered per phonetic transition (van Santen, 1997; Edgington, 1997; Sproat *et al.*, 1999). Another problem with such large databases is that it becomes harder for the speaker to maintain a consistent voice quality and pronunciation style. This problem was addressed by Stylianou (1999), who proposed to normalize the acoustic space of different recording sessions using an automatically derived filter.

As the phonetic and prosodic variability of the speech database increases, it also becomes more difficult to avoid acoustic mismatch between concatenated units. The solution to this problem is to expand the unit database even further, so that concatenations occur mostly in "safe" regions of a target utterance, for example in voiceless phonemes (Klabbers and Veldhuis, 2001).

1.3 Proposed Approach

To improve control of acoustic-prosodic variations in concatenative speech synthesis, two challenges must generally be addressed:

- A first challenge is to understand the effects of prosodic factors on a given acoustic variable, such as the fundamental frequency or pitch, and to predict variations of this variable for any input text.
- A second challenge is to modify acoustic units so that the predicted acoustic variations can be realized in a concatenated utterance.

In this thesis, acoustic-prosodic variations related to the *degree of articulation* of sonorant phonemes are investigated. Rather than collecting acoustic units in different prosodic contexts, we study the effects of prosodic factors on the degree of articulation of phonemes in natural speech, and develop a quantitative model to predict these variations for an utterance to be synthesized. Then, we modify the spectral shape of acoustic units, so that the desired degree of articulation is achieved.

The proposed approach relies critically on controlling the *spectral rate-of-change* of acoustic units. According to Lindblom (1963) and Moon and Lindblom (1994), the degree of articulation of sonorant phonemes, such as vowels, depends on three factors: (1) the phonetic context, (2) the phoneme duration, and (3) the spectral rate-of-change. The third factor is related to the speed of articulatory movements in the vocal tract, and reflects the articulation effort used by a speaker. In this work, the spectral rate-of-change is defined as the average rate-of-change of the resonant frequencies, or formants, at a given time point in sonorant speech.

The contextual model of articulation is illustrated in Figure 1.1, where one hypothetical formant trajectory of a vowel is shown. The difference between the ideal formant target

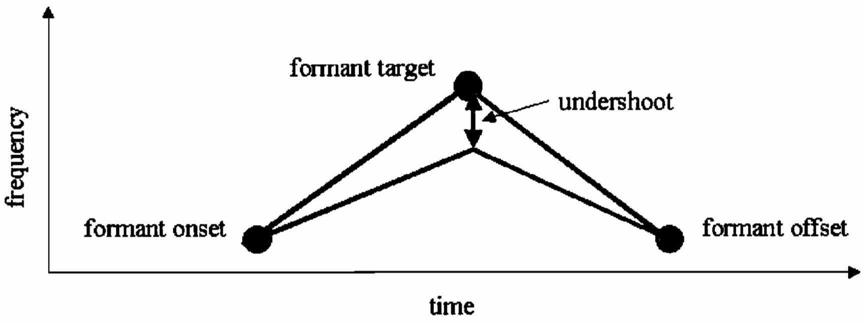

Figure 1.1: Contextual model of articulation: the amount of target undershoot depends on: (1) the phonetic context, (2) the phoneme duration, and (3) the spectral rate-of-change.

and the actual value reached at the vowel center is defined as the *target undershoot*. For a given phonetic context and given vowel duration, the amount of undershoot depends on the formant slope, i.e., on the spectral rate-of-change.

In Figure 1.1, the effect of the phonetic context on the degree of articulation is captured by the vertical distance between the formant onset and target values. If the target frequency is close to the onset frequency, undershoot is not likely to occur, because a low articulation effort, i.e., a small spectral rate-of-change, is sufficient to reach the vowel target (van Bergem, 1993). Hence, the contextual model contradicts the *centralization* view often held by speech researchers, according to which vowels tend to a "neutral" target in unstressed syllables, independent of the phonetic context (Lindblom, 1963; van Bergem, 1993).

The contextual model can be integrated in a concatenative synthesis system, by considering the spectral rate-of-change of phonemes as an independent parameter, in addition to phoneme identity and durations. Hence, two hypotheses are defended in this thesis:

1. the spectral rate-of-change of a particular phonetic transition is related in a predictable way to the prosodic context of a transition, and

2. modification of the spectral rate-of-change in concatenative synthesis enables control of the degree of articulation of acoustic units, and therefore can produce more

intelligible and more natural sounding synthetic speech.

The first hypothesis is motivated by reports in the literature that the spectral rate-of-change, or the speed of the articulators in the vocal tract, depends on the articulation *effort* used by a speaker (i.e., Kozhevnikov and Chistovich, 1965; Nelson, 1983; Flege, 1988; Moon and Lindblom, 1994). In turn, we expect that the speaker's articulation effort is determined by the prosodic context, i.e., by the relative prominence of phonemes, syllables, words, or phrases in an utterance, as well as more global characteristics related to the speaker's attitude or emotion (Lindblom, 1990). Hence, we hypothesize that a quantitative relation can be established between the spectral rate-of-change of acoustic units and prosodic factors, such as syllabic stress, pitch accent, word position, and speaking style.

The second hypothesis states that control of dynamic spectral features provides a key to improve the naturalness of concatenated speech. In formant synthesis, the phonetic quality of vowels can be controlled by redefining formant *targets* in specific phonetic and prosodic contexts (Klatt, 1987; Kohler, 1990). However, no clear view has emerged to predict the amount of target undershoot and the optimal formant shapes for different phonetic and prosodic contexts (e.g., Hertz, 1991; van Bergem, 1993). Moreover, parametric control of formant targets is typically not available in concatenative synthesis, due to problems with identifying and modifying formants in recorded speech. According to the contextual model of articulation, target undershoot does not reflect an intentional change in the articulatory target during human speech production, but it is a secondary effect of changes in timing and articulation effort. Therefore, we argue that the degree of articulation in concatenative synthesis should be controlled by modifying the spectral *dynamics* of acoustic units, in addition to timing. Our approach does not require explicit modeling of formant trajectories, and is motivated by the effects of prosodic context on the articulation effort in human speech.

The proposed approach also presents a solution to the problem of spectral mismatch between concatenated acoustic units, which produces unnatural spectral transitions or "glitches" in concatenated speech, often degrading the perceptual quality. Using the methods developed in this thesis, spectral mismatch can be reduced by specifying appropriate spectral dynamics at the concatenation points between units, and modifying the spectral shape of the acoustic units corresponding to the predicted dynamics.

1.4 Thesis Outline

The thesis consists of three parts, which were published or have been submitted as separate journal papers. Each chapter has its own introduction and provides a summary of the relevant background literature. Chapters 2 and 3 were part of the same paper, but are presented separately here for clarity.

In Chapter 2, a new technique is proposed to modify the spectral shape of acoustic units, while preserving the original perceptual quality. Previous attempts at modifying the spectral shape of speech units were often based on a decomposition of the speech signal into a filter modeling the transfer function of the vocal tract, and a residual approximating the excitation at the vocal chords. However, recombination of the residual with a modified transfer function introduces artifacts that decrease the perceptual quality of the modified speech. We describe a new technique for spectral modification, which is based on a sinusoidal + all-pole representation of speech, and avoids the shortcomings of modifications based on the source-filter decomposition.

In Chapter 3, the problem of concatenation mismatch between acoustic units is reviewed, and a method is presented to alleviate such mismatch. The method, called "unit fusion", is based on selecting two tiers of acoustic units in parallel, denoted as *fusion units* and *concatenation units.* The fusion units define the target spectral transitions at the join points between the concatenation units. The unit fusion method is further extended to control the articulation style of concatenated units when the speaking rate is increased.

In Chapter 4, the effects of prosodic factors on the spectral rate-of-change of phonetic transitions are analyzed for a balanced corpus of natural speech. The results show that the spectral rate-of-change increases with the linguistic prominence of a transition, i.e., in stressed syllables, in accented words, in sentence-medial vs. sentence-final words, and in clearly articulated speech. A numerical model is proposed to predict the spectral rate-of-change from prosodic factors.

In Chapter 5, the contextual model of articulation is integrated in a concatenative synthesis system. The target spectral rate-of-change of acoustic units is predicted based on the prosodic structure of utterances to be synthesized. Then, the spectral shape of the acoustic units is modified according to the predicted spectral rate-of-change, based on methods developed in Chapter 2. Experiments show that the proposed approach provides control over the degree of articulation of acoustic units, and improves the naturalness and intelligibility of concatenated speech in comparison to standard concatenation methods.

Chapter 2

Spectral Modification of Speech [1]

In this chapter, we review signal representations commonly used in concatenative speech synthesis, and we propose a signal processing technique to improve modification of the spectral shape of acoustic units. Current methods for spectral modification often rely on residual-excited linear prediction (RELP), which introduces artifacts in the resulting speech. We propose a new method based on a sinusoidal + all-pole representation of speech. This method enables high-quality spectral modifications of speech, while avoiding the shortcomings of RELP.

2.1 Pitch-Synchronous Analysis and Synthesis

Sonorant speech can be modeled as the output of a filter excited by a pulse-like source signal. The source signal corresponds to the excitation produced at the human glottis, while the filter models the resonances of the vocal tract (Fant, 1960). In sonorant speech, the excitation pulsates at a certain *fundamental frequency* or $F0$, corresponding to a perceived *pitch*. One cycle of the excitation waveform is called a *pitch period*. Usually, the start of a pitch period is defined as the moment of closure of the vocal chords, which is noticeable as an abrupt change in the excitation waveform (Childers and Ahn, 1995). When a speaker varies his or her pitch throughout an utterance, $F0$ changes relatively slowly from pitch period to pitch period, hence the excitation is said to be *quasi*-periodic.

[1] © IEEE. This chapter was published as part of Johan Wouters and Michael W. Macon: "Control of spectral dynamics in concatenative speech synthesis," and was reprinted, with permission, from *IEEE Transactions on Speech and Audio Processing* vol. 9, nr. 1, pp. 30–38, Jan. 2001. This work was supported by NSF Challenge grant CDA-9726363, NSF CAREER grant IIS-9875950, and by Intel Corporation.

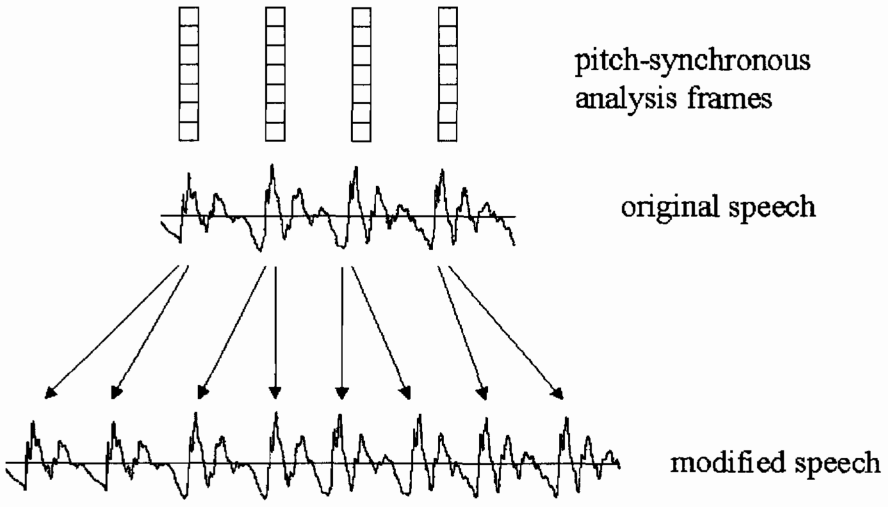

Figure 2.1: Illustration of pitch-synchronous analysis and duration modifcation. At each moment of glottal closure in the original speech waveform, an analysis frame representing two pitch periods is extracted. The duration of the original speech unit is doubled by assigning each analysis frame to two consecutive pitchmarks in the target waveform, and applying overlap-add.

In concatenative speech synthesis, it is desirable to modify certain characteristics of acoustic units, such as their duration, fundamental frequency, or aspects of their spectral shape. Most modification algorithms are based on a pitch-synchronous representation of the acoustic units, in order to separate effects of $F0$ on the speech acoustics from effects of the vocal tract. At each *pitchmark*, i.e., moment of glottal closure, speech is represented by a vector or *frame*. In time-domain modification methods, the frame consists simply of a waveform segment representing the two pitch periods centered around the pitch mark. In spectral-domain modification methods, the frame can contain a short-term spectral representation of the waveform, for example in the form of all-pole or sinusoidal parameters. These representations are explained in more detail below. If the order p of the spectral representation is constant, the sequence of pitch-synchronous frames defines a set of p parameter tracks, which we will refer to as *spectral trajectories*.

Pitchmarks can be computed from a speech signal directly, but this approach is sensitive to errors typically encountered in automatic pitch detection, such as pitch doubling or halving errors, and voiced/unvoiced misclassifications (Talkin, 1995). In our work, a *laryngograph* signal is recorded as part of the speech data collection process. This signal is a measure of the impedance variations at the larynx of the speaker, and can be processed more easily to compute the moments of glottal closure (Childers and Ahn, 1995).

To generate speech from a pitch-synchronous representation, a technique called *overlap-add* is used. Based on the desired duration and pitch of acoustic units, for example as predicted by a TTS system, target pitchmarks are computed. The target pitchmarks are associated with analysis pitchmarks in the original acoustic units. For example, if the duration of an original unit is doubled, this can be achieved by associating pairs of target pitchmarks to the same analysis pitchmark, as illustrated in Figure 2.1. In time-domain methods, the final speech waveform is produced by windowing the analysis frames and combining them using overlap-add. This method is known as time-domain pitch-synchronous overlap-add, or TD-PSOLA (Moulines and Charpentier, 1990). In spectral-domain methods, the spectral representation can first be modified, and is then converted back to a time-domain waveform. Finally, windowing and overlap-add are applied.

2.2 Residual-Excited LPC

Linear predictive coding (LPC) can be used to estimate an all-pole filter from a speech signal (Makhoul, 1975). If $S(\omega)$ represents the all-pole filter in the frequency-domain, the speech waveform can be passed through the inverse filter $1/S(\omega)$ to yield a *residual* signal, which approximates the excitation. In digital signal processing, $S(\omega)$ is implemented as an all-pole infinite impulse response (IIR) filter, while $1/S(\omega)$ represents an all-zero finite impulse response (FIR) filter. The inverse filter is applied pitch-synchronously, i.e., at each pitchmark, $S(\omega)$ is estimated and applied to a two period waveform segment, after which overlap-add yields the final residual signal.

A straightforward way to alter the spectral contents of a speech signal is to pitch-synchronously pass its residual through a target filter $S'(\omega)$, corresponding to the target

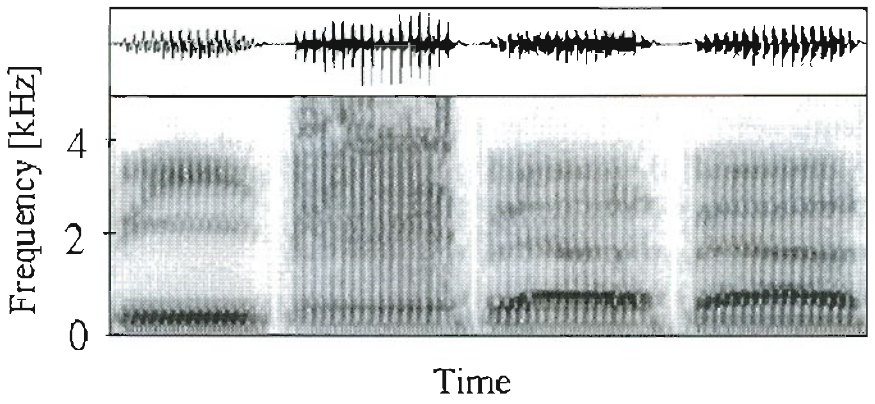

Figure 2.2: Effect of inverse filtering error on spectral transformation from /i:/ (as in "beet") to /æ/ (as in "bat"). From left to right: natural utterance of /i:/, residual of /i:/, transformation of /i:/ to /æ/ using RELP, natural utterance of /æ/. Large energy values are observed in the residual around the onset of F3 and below F2, leading to irregular formant structure in the modified vowel.

spectral shape. However, the LPC residual is not spectrally flat. It contains error in certain frequency regions, due to imperfect modeling of the original speech spectrum by the filter $S(\omega)$. The error in the residual can interfere with the formant structure of the target spectrum $S'(\omega)$. In particular, Makhoul (1975) showed that the all-pole model error is small around resonant frequencies and large in the spectral valleys. Therefore, when an LPC residual is passed through a target filter $S'(\omega)$, relatively large errors in spectral valleys of the original filter $S(\omega)$ may be amplified at resonant frequencies in $S'(\omega)$. This leads to irregular formant structure in the modified speech, as illustrated in Figure 2.2.

2.3 Sinusoidal Representation

To overcome the limitations of spectral modification based on RELP, more flexible control over the spectral components of the speech signal is needed. Sinusoidal modeling has been demonstrated to enable high-quality speech coding (McAulay and Quatieri, 1986), pitch and time scale modifications in synthesis (Macon, 1996; Laroche *et al.*, 1993), and voice conversion (Stylianou, 1996; Stylianou *et al.*, 1995a). Here, we explore the use of this model for spectral modification of speech.

The basis of the sinusoidal model is a decomposition of the speech signal $s[n]$ into a sum of time-varying sinewaves oscillating at multiples of the fundamental frequency ω_0:

$$\hat{s}[n] = \sum_{k=-L}^{L} a_k \exp(jk\omega_0 n) = \sum_{k=-L}^{L} A_k \exp(j(k\omega_0 n + \phi_k)), \qquad (2.1)$$

where L is the number of harmonics, and $a_k = A_k \exp(j\phi_k)$ is the complex amplitude of the kth harmonic.

The sinusoidal parameters $\{a_k\}$ are estimated by solving a complex linear regression, as described by Laroche *et al.* (1993) and Stylianou (1996, 2001). Sinusoidal *synthesis* can be achieved by evaluating Equation (2.1) for every speech sample. The sinusoidal parameters are either interpolated between analysis points or frame-based overlap-add is used in the time-domain (Stylianou *et al.*, 1995b; George and Smith, 1997; Stylianou, 2001).

Modifications in the spectral structure of $s[n]$ can be achieved by defining a transformation from the complex amplitudes $\{a_k\}$ to a new set $\{a'_k\}$. However, the dimensionality of the sinusoidal representation can be as high as that of the sampled speech waveform, making direct transformations of $\{a_k\}$ impractical.

2.4 Sinusoidal + All-Pole Modeling

An advantage of spectral modification based on the source-filter model is that the LPC filter describes the spectral shape with relatively few coefficients, typically on the order of 10 to 20 coefficients per frame. The dominant poles of the LPC filter are also closely related to formant frequencies and bandwidths, which is advantageous for specifying spectral modifications as in Sections 3.2 and 3.3. Therefore, we approximate the sinusoidal parameters $\{a_k\}$ by an all-pole model $S(\omega)$.

Estimation of all-pole parameters based on a discrete power spectrum is a well-known technique in speech processing (Makhoul, 1975; Hermansky, 1990), and was initially applied to the sinusoidal model by McAulay (1984); McAulay and Quatieri (1995). In our system, the power spectrum $|a_k|^2$ is first mel-warped to improve the resolution of the all-pole model towards lower frequencies, corresponding to the frequency resolution of human

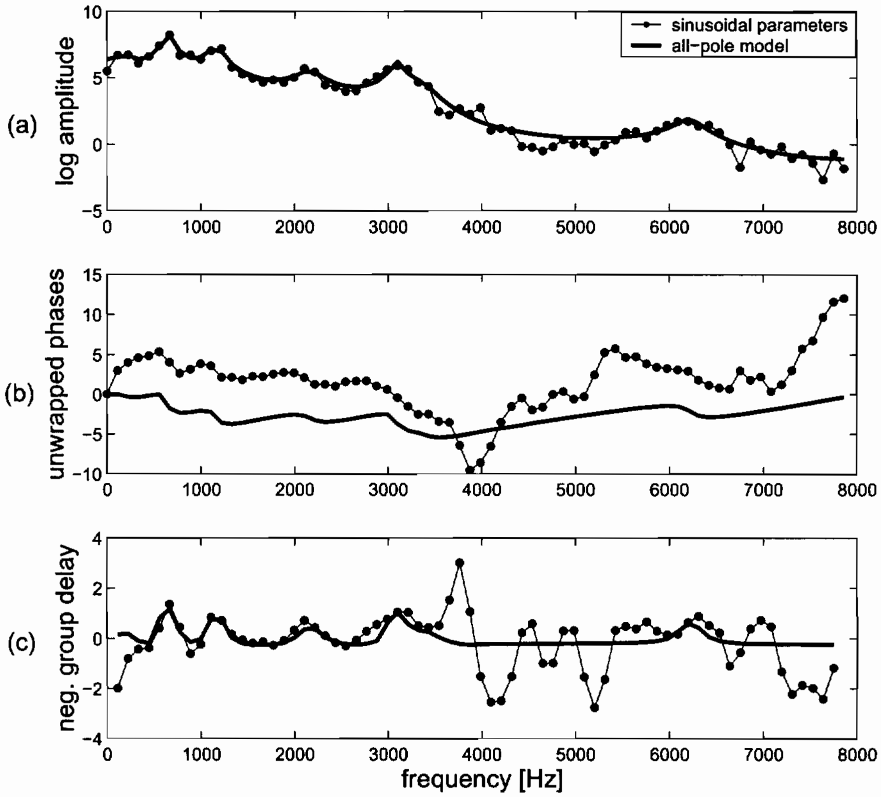

Figure 2.3: Comparison between sinusoidal parameters $\{A_k, \phi_k\}$ and all-pole model $S(\omega)$: (a) log amplitude spectra, (b) unwrapped phase spectrum, (c) group delay spectrum.

hearing. To reduce bias of the all-pole spectrum towards the harmonic frequencies, the power spectrum is upsampled, using cubic interpolation in the log domain (Hermansky *et al.*, 1984). Then, correlation coefficients are obtained by an Inverse Discrete Fourier Transform (IDFT) of the power spectrum. Application of the Levinson-Durbin algorithm to the correlation coefficients yields the all-pole (LPC) parameters. We refer to the joint representation of $\{a_k\}$ and $S(\omega)$ as the *sinusoidal + all-pole* model.

In Figure 2.3, the smooth all-pole spectral envelope can be compared with the sinusoidal parameters for an example vowel segment. An 18th order all-pole model was used, for speech sampled at 16 kHz. The all-pole envelope approximates the sinusoidal magnitudes A_k quite well, especially in the formant regions. In the first derivative of the phase, or group delay, the phase relationships between successive frequency components can be investigated. The negative group delay of the speech spectrum reflects the formant structure (Murthy and Yegnanarayana, 1991), as can be verified in Figure 2.3(c). Therefore, the all-pole phases in Figure 2.3 seem to approximate the sinusoidal phase characteristics reasonably well, at least in the formant regions. The large discrepancies between the unwrapped phase spectra in Figure 2.3(b) are due to accumulated mismatch between the all-pole model phase and the phase of the first sinusoidal components, as well as the components above the third formant.

When sinusoidal synthesis is performed with the complex amplitudes $S(k\omega_0)$, as opposed to the "exact" sinusoidal parameters $\{a_k\}$, artificial-sounding speech is produced, similar in quality to an impulse-excited LPC vocoder. This is because the all-pole envelope $S(\omega)$ models broad-scale information about the speech spectrum, such as formant frequencies and bandwidths, but it does not capture some perceptually important details of the amplitudes $\{a_k\}$. In the next section, we present a method that utilizes the smooth envelope $S(\omega)$ to *transform* $\{a_k\}$ to fit a target envelope $S'(\omega)$, without requiring that the all-pole model describes all the spectral detail in the sinusoidal amplitudes.

2.5 Modification of the sinusoidal parameters

Based on the sinusoidal + all-pole representation, we describe a spectral modification algorithm that allows modification of the speech spectrum to approximate a target all-pole model. The method maintains high speech quality and avoids undesired interaction between the target filter and the residual, as occurs in RELP. The algorithm is illustrated in Figures 2.4 and 2.5.

We assume that a speech segment has been analyzed such that sinusoidal parameters $\{a_k\}$ and an all-pole model $S(\omega)$ approximating $\{a_k\}$ are known. Also, we assume that a target all-pole spectrum $S'(\omega)$ has been established, for example using the methods presented in Sections 3.2 and 3.3. In Figure 2.4, the magnitudes of $\{a_k\}$, $S(\omega)$ and $S'(\omega)$ are shown for an example vowel. The second formant has shifted to a higher frequency in $S'(\omega)$, while the first and third formant have moved to a somewhat lower position. The

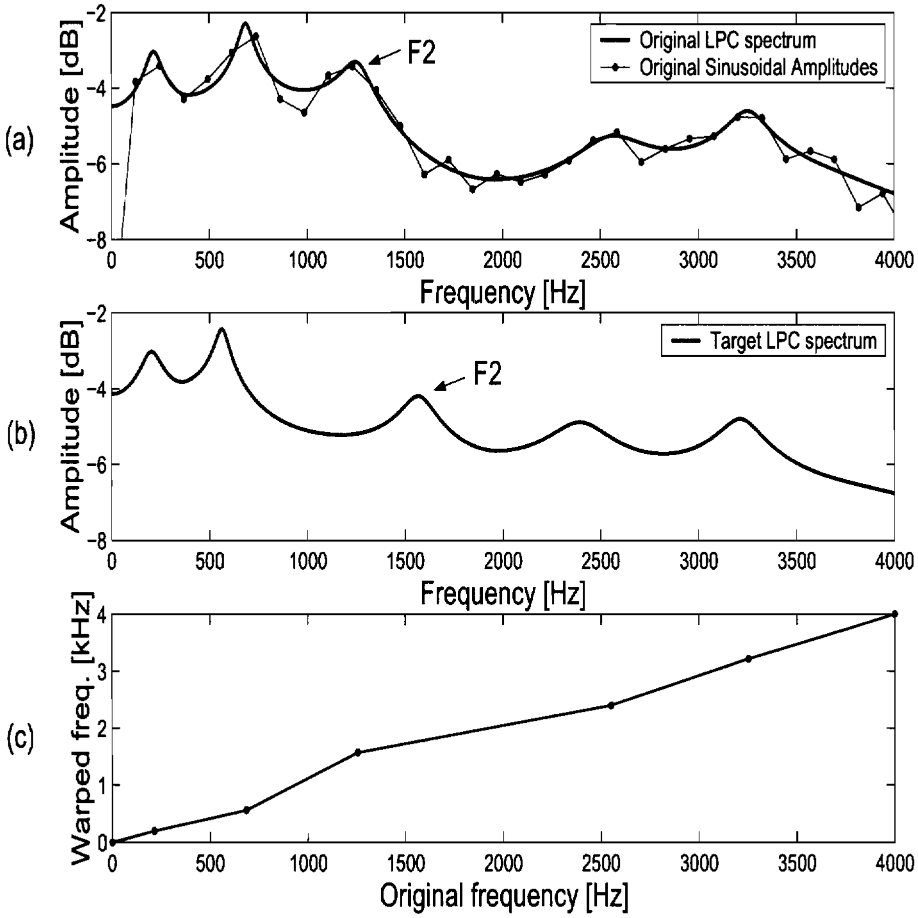

Figure 2.4: Input for spectral modification algorithm: magnitude spectra are shown for (a) original sinusoidal parameters $\{a_k\}$ and all-pole model $S(\omega)$, and (b) target all-pole model $S'(\omega)$. In (c), the piece-wise linear frequency warping is shown which maps dominant poles of $S(\omega)$ to dominant poles of $S'(\omega)$.

first peak in the spectrum corresponds to a characteristic of the excitation (i.e. "glottal formant") as opposed to a resonance of the vocal tract. The challenge is to find sinusoidal parameters $\{a'_k\}$ that fit the new spectral envelope, but also preserve the (unmodeled) spectral detail observed in $\{a_k\}$.

First, a frequency warping function $f : \omega \to \omega'$ is determined, which maps the dominant poles of $S(\omega)$ to the dominant poles of $S'(\omega)$. A piece-wise linear warping can be used, as illustrated in Figure 2.4(c). Estimation of $f(\omega)$ is aided by the fact that $S(\omega)$ and $S'(\omega)$ describe similar speech sounds, and that $S'(\omega)$ may have been obtained by manipulating $S(\omega)$, as described in Sections 3.2 and 3.3.

Then, $S'(\omega)$ is non-uniformly sampled at frequency points $\omega'_k = f(k\omega_0)$, where $f(\omega)$ is the frequency warping function defined earlier. This yields

$$b_k = S'(\omega'_k) = S'(f(k\omega_0)), \tag{2.2}$$

which is illustrated in Figure 2.5(a). Each frequency point ω'_k is located in the same position with respect to the dominant poles of $S'(\omega)$ as the harmonic $k\omega_0$ is located with respect to the dominant poles of $S(\omega)$.

In the next step, spectral detail encoded in the original parameters $\{a_k\}$ is *transferred to corresponding regions* of the target spectrum. This is expressed by setting

$$b'_k = \frac{a_k}{S(k\omega_0)} b_k. \tag{2.3}$$

In this operation, the residual error between a_k and $S(k\omega_0)$, defined as $a_k/S(k\omega_0)$, is first frequency-warped according to $f(\omega)$, and is then combined with component b_k of the non-uniformly sampled target envelope. Consequently, similar patterns of all-pole model error are found in the spectral valleys of $S(\omega)$ and in the valleys of $S'(\omega)$, and similarly around the spectral peaks. This can be verified in Figures 2.4(a) and 2.5(b). Our approach therefore avoids the problems of the RELP method, in which model error in the valleys of the original speech can interfere with the formant regions of the target speech. Furthermore, in frequency regions where $S(\omega) \approx S'(\omega)$, the original sinusoidal amplitudes are preserved.

Equations (2.2) and (2.3) represent complex variables, hence the modifications apply to both the sinusoidal amplitudes A_k and the phases ϕ_k. We have observed that it is important

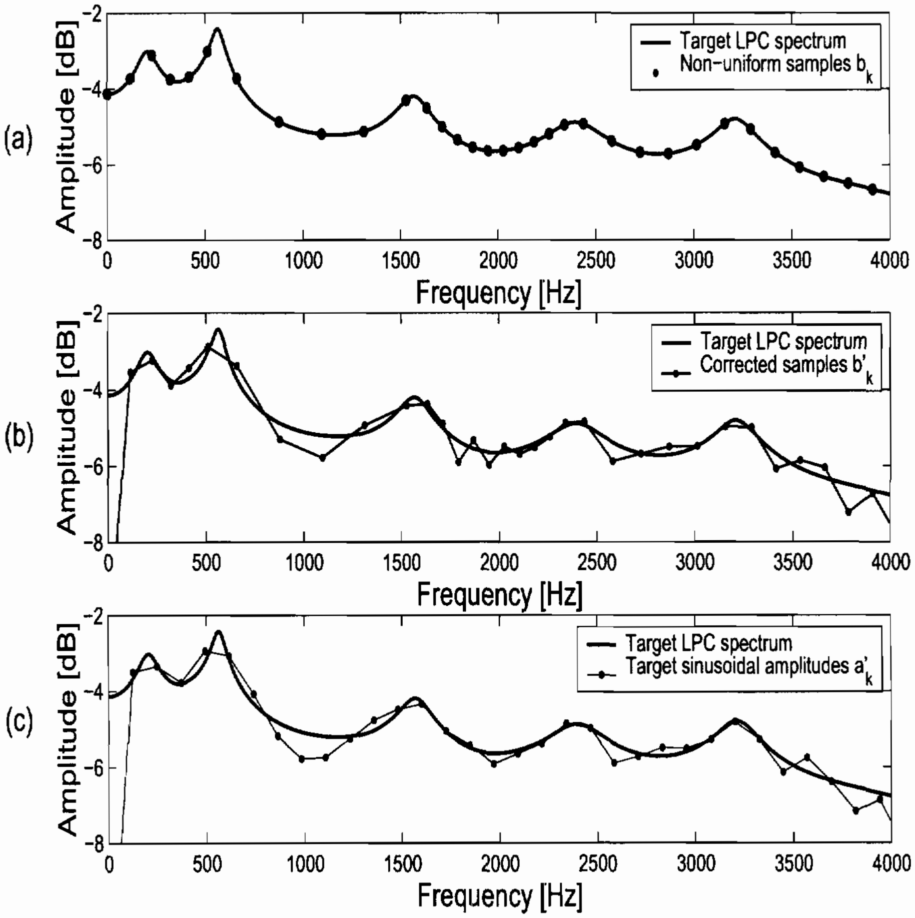

Figure 2.5: Computation of target sinusoidal parameters $\{a'_k\}$: magnitude spectra are shown for (a) non-uniform sampling of target all-pole spectrum, (b) transfer of original all-pole model error $a_k/S(k\omega_0)$, (c) target sinusoidal parameters resampled at $k\omega'_0$. The pattern of all-pole error in (c) is similar to Figure 2.4(a).

to maintain the relationship between resonant peaks in the amplitude spectrum and peaks in the group delay spectrum, as illustrated in Figure 2.3. In Equations (2.2) and (2.3), phase characteristics that are not modeled by $S(\omega)$ are transferred to equivalent spectral regions in $S'(\omega)$, in conjunction with amplitude modifications. Informal listening tests showed no perceptual difference between modifying the phases directly as in Equations (2.2) and (2.3), or modifying the group delay spectrum. On the other hand, when the sinusoidal phases ϕ_k were not modified at all (i.e. only magnitudes were considered in (2.2) and (2.3)), a harsh quality was introduced in the speech signal. The relationship between amplitude and phase modifications is an aspect of our ongoing research.

The final step of the algorithm is to resample $\{b'_k\}$, as the new sinusoidal parameters $\{a'_k\}$ should coincide with harmonics of the target fundamental frequency ω'_0 while $\{b'_k\}$ is defined at the frequencies $f(k\omega_0)$. To obtain $\{a'_k\}$, we employ cubic interpolation on the values $\log(|b'_k|)$ and linear interpolation on the unwrapped phases $\angle b'_k$. The result is shown in Figure 2.5(c) for the example magnitude spectrum. In comparison with Figure 2.4(a), the new parameters $\{a'_k\}$ are not only frequency-warped with respect to $\{a_k\}$, but the amplitudes have changed according to the shape differences between $S(\omega)$ and $S'(\omega)$.

2.6 Evaluation

In Wouters and Macon (2000), we evaluated an earlier version of the spectral modification method, which was based on the sinusoidal + all-pole model but did not add the warped spectral residual in the formant regions, i.e., between 300 Hz and 4 kHz. We recorded a male and a female speaker who repeated the sentence "Please say b_t again", with four different target words: *beat*, *bat*, *bought* and *boot*. We then generated 12 new sentences per speaker. First, the vowel in each target word was replaced by a *different* vowel, transplanted from another target word. Next, the transplanted vowels were spectrally modified to approximate the vowel they replaced.

We compared the sinusoidal + all-pole method with spectral modifcation based on inverse filtering. An 18th-order all-pole filter was used in both methods to transplant the spectral characteristics of the target vowel to the original vowel. The modifications

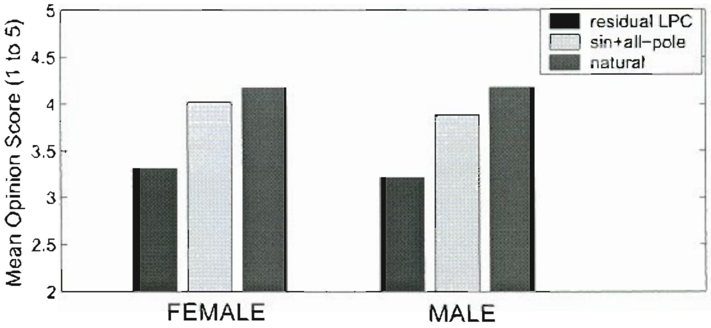

Figure 2.6: Perceptual quality scores. The scores are ranked per speaker according to the modification method used. From left to right: residual excited LPC method, sinusoidal + all-pole method, and without modification.

were performed pitch-synchronously and the waveform was reconstructed using overlap-add. The all-pole filter for inverse filtering was estimated over three pitch periods using the modified covariance method (the average F_0 is 200 Hz for the female speaker and 120 Hz for the male speaker). The duration and pitch of the modified vowel were adjusted by adopting the pitch marks from the target and dropping or duplicating analysis frames.

We presented the original and the modified sentences to 20 listeners, and asked them to identify the target word in each sentence and to rate the perceptual quality on a five-point scale. Prior to the test, the listeners ran through a set of examples so they had an idea of the best and worst cases to expect in terms of quality. (The training set was generated for a different sentence "Please say s_t again.")

All target words were correctly identified by the listeners. The perceptual quality scores are summarized in Figure 2.6. For the female voice, the average score was 3.31 for the residual excited LPC method and 4.02 for the sinusoidal + all-pole method, compared to 4.17 for the unmodified sentences. For the male voice, the average quality score was 3.21 for the residual excited LPC method and 3.88 for the sinusoidal + all-pole method, compared to 4.17 for the unmodified sentences. A clear preference is shown for the sinusoidal + all-pole method, and this method causes little degradation compared to unmodified speech.

Since natural targets for each vowel modification were available, we were also able

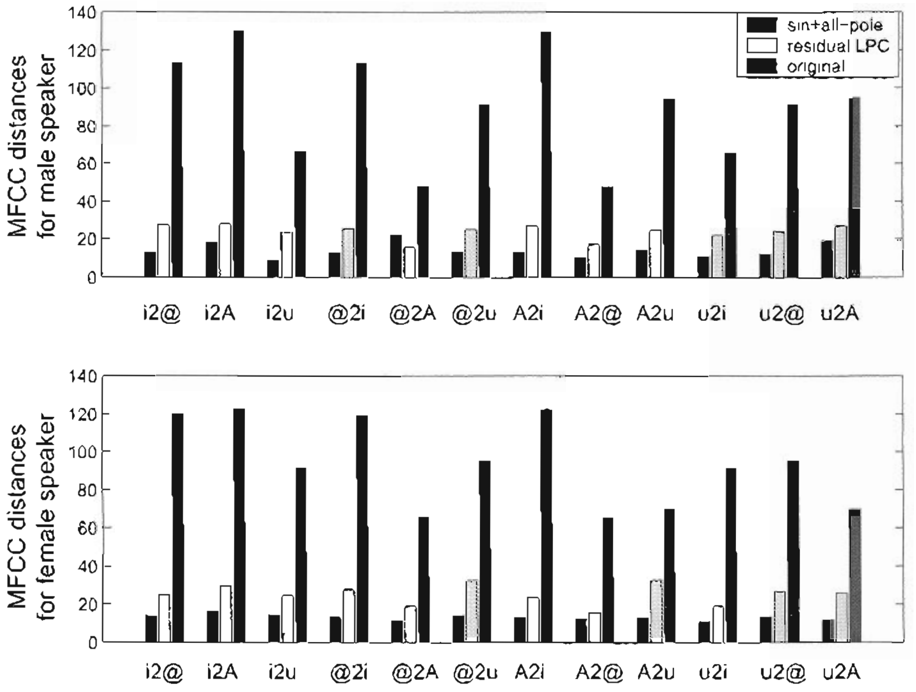

Figure 2.7: Mel cepstral distances between modified vowels and target vowels. For each warping from vowel X to vowel Y (labeled 'X2Y'), the distances are ranked according to the method used: sinusoidal + all-pole method, residual excited LPC, and without modification.

to compute objective distances between the modified vowels and the targets. Thus we could verify our hypothesis that the sinusoidal + all-pole method leads to spectra that are closer to the target than those produced by the inverse filtering method. This is shown in Figure 2.7. Mel cepstral distances were computed every 5 milliseconds with a window size of 30 milliseconds and averaged over the number of frames in the vowel. The distances between the different natural vowels are shown for comparison.

In the next chapters, we discuss methods to specify target all-pole spectra $S'(\omega)$ in order to reduce concatenation mismatch, and to control the degree of articulation of sonorant speech units. Resynthesis of the acoustic units corresponding to the target spectrum will be based on the sinusoidal + all-pole representation as described in the previous sections.

Chapter 3

Control of Spectral Dynamics in Concatenative Speech Synthesis [1]

Current speech synthesis methods based on the concatenation of waveform units can produce highly intelligible speech capturing the identity of a particular speaker. However, the quality of concatenated speech often suffers from discontinuities between the acoustic units, due to contextual differences and variations in speaking style across the database. In this chapter, we present methods to spectrally modify speech units in a concatenative synthesizer to correspond more closely to the acoustic transitions observed in natural speech. First, a technique called "unit fusion" is proposed to reduce spectral mismatch between units. In addition to concatenation units, a second, independent tier of units is selected that defines the desired spectral dynamics at concatenation points. Both unit tiers are "fused" to obtain natural transitions throughout the synthesized utterance. The unit fusion method is further extended to control the perceived degree of articulation of concatenated units.

3.1 Introduction

Most current text-to-speech systems generate speech by concatenating recorded waveforms. A recent trend is to collect very large databases of fluent speech and to select an optimal

[1] © IEEE. This chapter was published as part of Johan Wouters and Michael W. Macon: "Control of spectral dynamics in concatenative speech synthesis," and was reprinted, with permission, from *IEEE Transactions on Speech and Audio Processing* vol. 9, nr. 1, pp. 30–38, Jan. 2001. This work was supported by NSF Challenge grant CDA-9726363, NSF CAREER grant IIS-9875950, and by Intel Corporation.

sequence of acoustic units at run-time to synthesize a particular utterance. The advantage of this approach is that problems with joining or modifying certain speech units can be avoided if suitable sentence fragments are found in the database (Hunt and Black, 1996; Donovan, 1998; Beutnagel *et al.*, 1998; Breen and Jackson, 1998a). However, for synthesis from unrestricted text, the challenge to collect all required variations of acoustic units in a finite database remains unanswered. As a result, occasional spectral mismatch between concatenated units has to be tolerated, as do artificial changes in speaking style within an utterance.

We propose methods for improving the quality of acoustic transitions in synthetic speech by modifying the spectral shape of concatenated units. Two challenges must be addressed to enable high-quality spectral transformations: (1) changes in the spectral shape must be specified which will improve synthesis quality, for example by removing spectral discontinuities, (2) a signal processing method must be established to resynthesize units with a modified spectral shape, while maintaining the original speech quality. Both challenges are addressed in this chapter.

In Section 3.2, we present a method called *unit fusion* which is designed to reduce spectral discontinuities between concatenated units. The method is based on selecting two independent tiers of speech units, denoted as *concatenation units* and *fusion units*. The concatenation units represent initial spectral trajectories for an utterance, while the fusion units specify the desired transitions at the concatenation points. The information from the two tiers is "fused" in order to obtain natural transitions throughout the synthetic utterance.

The fusion process relies on a cost function that is minimized to yield smoothed spectral trajectories. This cost can be further exploited to control the degree of articulation in concatenated speech. In Section 3.3 we review a model of human speech articulation developed by Lindblom (1990). According to this model, increased articulation effort corresponds with faster movements of the articulators in the human vocal tract, and hence faster transitions in the acoustic signal. We describe results of a method for modifying the dynamic characteristics of speech in correspondence with this model.

3.2 Removing Concatenation Discontinuities

An important problem to be addressed in concatenative synthesis is the occurrence of sudden spectral changes due to mismatch between concatenated units. While rapid transitions are common in natural speech (e.g., plosives), spectral discontinuities between concatenated units in sonorant speech have a disturbing perceptual effect. In this section, we describe a technique for reducing concatenation mismatch, by *merging* spectral information from independently selected speech units.

3.2.1 Background

Several researchers have developed approaches to minimize spectral mismatch in concatenated speech. For example, systems that select acoustic units at run-time often incorporate a "concatenation cost," based on a measure of spectral discontinuity, and search for units that minimize this cost (Hunt and Black, 1996; Conkie and Isard, 1997; Donovan, 1998; Beutnagel *et al.*, 1998). However, such methods can be successful only if acceptable concatenation points exist between the segments in the database. Furthermore, the selection may be suboptimal since the acoustic distance measures that are commonly used have only moderate correlation with human judgements of acoustic distortions (Wouters and Macon, 1998; Klabbers and Veldhuis, 1998).

Other techniques have been proposed to mitigate the effects of concatenation artifacts by modifying the spectral characteristics of speech. Most approaches are based either on waveform interpolation of pitch periods or on smoothing of LPC-derived parameters (Giménez *et al.*, 1994; Dutoit and Leich, 1994). In either approach, the region of interpolation is set by rule. The waveform interpolation technique is computationally inexpensive, but the speech spectra are merely cross-faded around the concatenation points, which can lead to spectral patterns not observed in natural speech (Dutoit and Leich, 1994). Certain parametrizations of the LPC coefficients, such as *line spectral frequencies* (LSF) allow smoothing that is more similar to formant movements (Dutoit and Leich, 1994; Paliwal, 1995; Yong, 1994). Line spectral frequencies are automatically derived from the LPC coefficients, and define trajectories akin to formant trajectories. Hence, they provide an elegant way to manipulate the spectral structure associated with a sequence of LPC filters.

In speech recognition, Hermansky and Morgan (1994) have proposed to filter the power spectrum of speech utterances, according to the time-integration properties of human hearing. Their technique, called RASTA, also has potential to smooth concatenation mismatches in speech synthesis. However, application of RASTA corresponds to a cross-fading of the spectral characteristics around concatenation points, similar to applying linear interpolation in the waveform domain (Dutoit and Leich, 1994). An additional problem is to reconstruct speech from the modified power spectrum, without introducing perceptually disturbing artifacts (Hermansky and Morgan, 1994; Takano and Abe, 1999).

Recently, Plumpe *et al.* (1998) proposed a technique in which line spectral frequencies and their first derivatives are characterized by Hidden Markov Models (HMM) trained on a large single speaker database (Plumpe *et al.*, 1998). After concatenating the units selected for a particular utterance, the corresponding LSF trajectories are each modified by minimizing the following error criterion with respect to the new LSF's x_i:

$$E = \sum_{i=1}^{N} \frac{(x_i - f_i)^2}{\sigma_i^2} + D \sum_{i=1}^{N-1} \frac{(\Delta x_i - \Delta f_i)^2}{\Delta\sigma_i^2}, \tag{3.1}$$

where f_i represents the original LSF trajectory at a time point i, Δf_i is the desired time-derivative (or "delta") at point i, Δx_i is the time-derivative of x_i, σ_i^2 and $\Delta\sigma_i^2$ are the variances of f_i and Δf_i, respectively, and N is the total number of time points in the trajectory. The criterion expresses a trade-off between (a) keeping x_i close to the original values f_i, and (b) forcing Δx_i to obey the constraints described by Δf_i. A weighting factor D determines the relative importance of each term.

The target values for Δf_i in Equation (3.1) are obtained during an HMM training process as the average time-derivative of LSF parameters extracted from acoustic segments clustered on each HMM state. However, since the statistics are computed from different clusters of speech data for each selected acoustic unit, the dynamic constraints Δf_i are not necessarily consistent across concatenation points. Plumpe *et al.* also reported problems due to minimizing the criterion in Equation (3.1) for each LSF trajectory independently. Since the distances between adjacent LSF coefficients are related to the pole-bandwidths of the corresponding all-pole filter, this can lead to spectrally smeared formants or to extremely sharp spectral peaks. Furthermore, in Plumpe *et al.* (1998), the output speech signal was regenerated using residual-excited LPC, which introduces artifacts as discussed in Section 2.

3.2.2 Unit Fusion

We propose a new approach to reduce concatenation mismatch which is a generalization of search-based unit selection as formulated by Hunt and Black (1996) and others. In the proposed approach, synthesis is performed by combining information from *two* tiers of speech units, denoted *concatenation units* and *fusion units.* The concatenation units specify initial estimates of the spectral trajectories for an utterance, while the fusion units characterize the spectral dynamics at the join points between concatenation units. These two unit tiers are "fused" during synthesis to obtain natural spectral transitions throughout the synthesized speech.

The fusion units are selected independently for each concatenation point by minimizing a linguistically motivated target cost. This cost takes into account phonetic and prosodic differences between the target utterance and the units available in the database. For example, if a concatenation point occurs at the center of a phoneme p, a fusion unit will be selected which represents p and whose phonetic context most closely matches the identity or place of articulation of the phonemes surrounding p in the target utterance. If multiple candidate units are found, further selection is based on the location of syllabic boundaries and the difference in duration with the target phoneme.

The fusion process is illustrated in Figure 3.1. Both concatenation and fusion units are represented by line spectral frequencies, based on the sinusoidal + all-pole representation discussed in Section 2. The new, smoothed, LSF trajectories are obtained by imposing dynamic constraints Δf_i on the LSF trajectories of the concatenation units. These dynamic constraints are computed from the time-derivatives of both the concatenation units and the fusion units, using the interpolation function α:

$$\Delta f_i = \alpha_i \Delta f_{k(i)}^{fusion} + (1 - \alpha_i) \Delta f_{l(i)}^{concat}, \tag{3.2}$$

where $k(i)$ and $l(i)$ are warping functions that map time frames in the target utterance

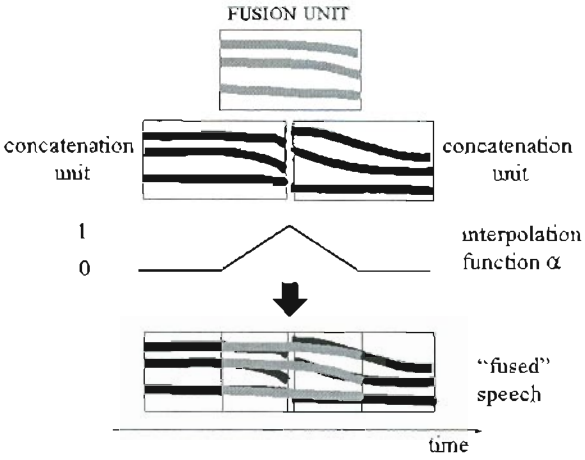

Figure 3.1: Illustration of unit fusion. The *time-derivatives* of the fusion unit and the concatenation units are interpolated using α, and the resulting spectral constraints are applied to the concatenation units using Equation (3.3) (see text).

to frames in the fusion and concatenation units, respectively. The interpolation function α_i reaches 1 at each concatenation point, which corresponds with the center of a fusion unit; it reaches 0 at the boundaries of the fusion unit, corresponding with a point inside the concatenation units. Hence, the fusion unit governs the spectral dynamics at the concatenation point, while the dynamics of the concatenation units are in force further away from the point of concatenation.

The constraints Δf_i can be imposed on the LSF trajectories using the cost function introduced in Equation (3.1). We extend this cost function with a third term in order to maintain appropriate distances between adjacent LSF trajectories. The cost function is minimized with respect to the parameters $x_{i,j}$, representing the new LSF trajectories:

$$\begin{aligned} E &= \sum_{i=1}^{N}\sum_{j=1}^{M}(x_{i,j} - f_{i,j})^2 \\ &+ \sum_{i=1}^{N-1}\sum_{j=1}^{M} D_1(i,j)(\Delta^i x_{i,j} - \Delta^i f_{i,j})^2 \end{aligned}$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{M-1} D_2(i,j)(\Delta^j x_{i,j} - \Delta^j f_{i,j})^2. \tag{3.3}$$

where $f_{i,j}$ represents an initial LSF value at time i in trajectory j, $\Delta^i f_{i,j}$ is the desired time-derivative at (i,j), and $\Delta^j f_{i,j}$ is the desired LSF distance, normally set to $(f_{i,j+1} - f_{i,j})$. Similarly, $\Delta^i x_{i,j}$ is the time-derivative of $x_{i,j}$ and $\Delta^j x_{i,j}$ represents the distance between adjacent LSF coefficients. N is the number of time points in each LSF trajectory and M is the number of trajectories, corresponding with the LPC model order.

The error criterion cannot be minimized by considering the LSF trajectories independently, as was done for Equation (3.1). If $\Delta^i x_{i,j} = x_{i+1,j} - x_{i,j}$ and $\Delta^j x_{i,j} = x_{i,j+1} - x_{i,j}$, then E becomes quadratic in the variables $x_{i,j}$. Setting $dE/dx_{i,j} = 0$ for each (i,j) defines a set of linear equations in $x_{i,j}$. This can be reformulated as a linear regression problem $A\mathbf{x} = \mathbf{b}$, where $\mathbf{x}$ represents the columns of $x_{i,j}$, one placed under the other. The linear regression problem is solved using standard methods as $\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$.

Two weighting coefficients determine the importance of the various constraints in Equation (3.3). If D_1 is increased, the time-derivative constraints will be imposed more heavily. If D_2 is increased, the distances between LSF pairs will be better preserved. In our experiments, D_1 was set to an empirically determined value of 20. $D_2(i,j)$ is made proportional to the inverse square of $\Delta^j f_{i,j}$, so that the distance between close LSF's is tightly controlled, while widely spaced LSF's can change more freely. Hence, the bandwidths of dominant poles are maintained and the risk of spurious peaks due to close placement of LSF coefficients is reduced.

3.2.3 Experimental Results

Twenty-five sentences were generated once using simple pitch-synchronous concatenation and once using the proposed fusion method. The sentences had been randomly chosen from a database read by the same speaker that recorded the unit databases. Intonation and phoneme durations were extracted from the natural sentences and imposed on the concatenated diphone units. In both synthesis techniques, TD-PSOLA (Moulines and Charpentier, 1990) was used to control pitch and durations of the units.

The first system is a time-domain method and no further modification of the concatenated diphones takes place. In the second system, PSOLA modifications were followed by a spectral modification step implementing the proposed unit fusion. The log energy contours of the concatenation units were also modified using information from the fusion units, by adding an extra trajectory to Equation (3.3). The spectrally modified units were regenerated using the sinusoidal + all-pole method described in Section 2, in order to avoid the artifacts involved with residual-excited LPC.

A comparative mean opinion score (CMOS) test was performed based on the guidelines in ITU P.800 (1996) to compare the performance of the time-domain method to the unit fusion method. Fifteen subjects listened to pairs of utterances and were asked to rate the quality of utterance "A" relative to "B". The utterances A and B corresponded to the same sentence generated by the two systems and were randomized per listener, both within and among pairs. A seven-point perceptual scale was used, as shown in Table 3.1. The subjects listened to five examples prior to taking the test. These examples had been chosen at random and were the same for all listeners.

Table 3.1: Perceptual scale for Comparison Mean Opinion Score (CMOS).

3.	Much Better
2.	Better
1.	Slightly Better
0.	About The Same
-1.	Slightly Worse
-2.	Worse
-3.	Much Worse

The perceptual score averaged over all the listeners and sentence pairs was 0.8 in favor of the unit fusion method ($p \ll 0.001$, two-tailed t-test). The average score per listener varied from 0.0 to 1.9, while the average per sentence pair varied from -0.4 to 1.3.

We then compared the unit fusion technique to a system in which linear smoothing was applied at the concatenation points between units. The algorithm described by Dutoit and Leich (1994) was used, in which the LSF differences at a concatenation point are spread

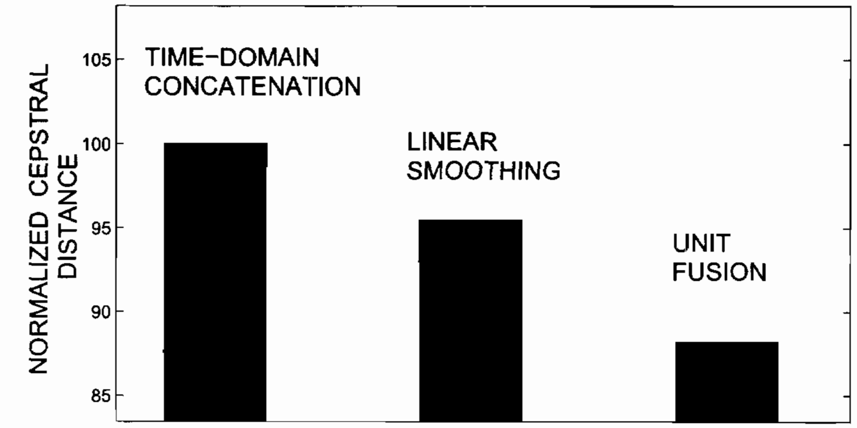

Figure 3.2: Normalized cepstral distance between natural utterances and synthetic speech generated by (1) time-domain concatenation, (2) linear smoothing, and (3) unit fusion.

linearly over a specified region. The interpolation region extended a maximum of 100 milliseconds to either side of the join point, limited by the duration of the smoothed phoneme. Linear smoothing was also applied to the log energy trajectory of the concatenated units. The signal was regenerated using the sinusoidal + all-pole method described in Section 2.

We observed that the linear technique degraded the phonetic quality of certain diphthongs and liquids when compared with the results of the unit fusion method. However, a perceptual evaluation comparing linear smoothing and unit fusion (using the same setup and listeners as described above) yielded an average score of 0.15, indicating a small preference for linear smoothing ($p = 0.007$, two-tailed t-test). We concluded that for the given test sentences, the phonetic degradations due to linear interpolation were not perceptually salient.

To evaluate the unit fusion technique using an objective measure, we computed spectral distances between the naturally spoken sentences and the synthetic utterances generated using (1) time-domain concatenation, (2) linear smoothing, and (3) unit fusion. Mel-warped cepstral distances were computed in sonorant speech regions using the all-pole representation described in Section 2. The normalized distances are shown in Figure 3.2. Compared with the time-domain concatenation method, linear interpolation reduces the

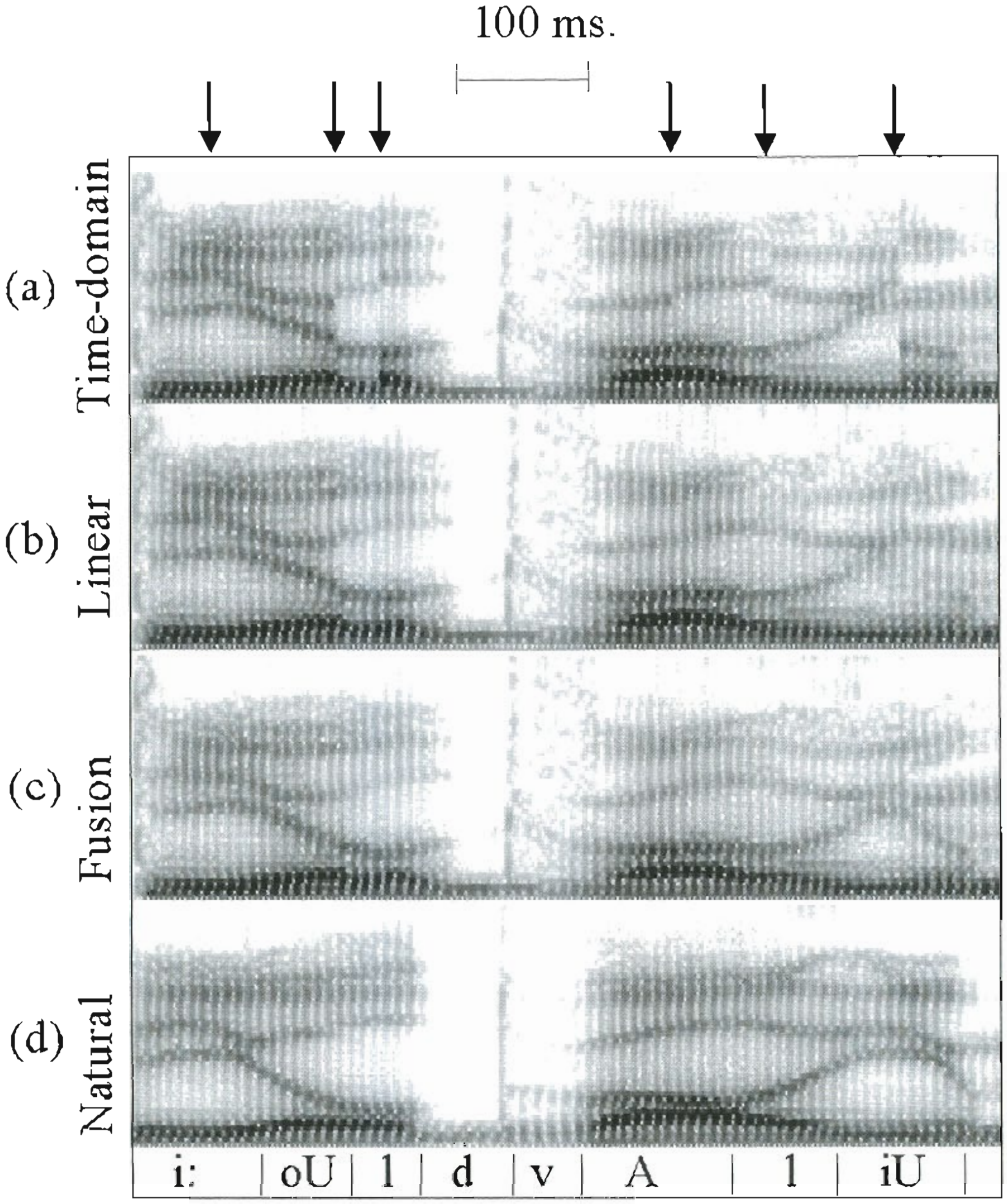

Figure 3.3: Spectrograms for "|dust|y old volu[me]" computed from (a) time-domain concatenation of units, (b) synthetic speech after linear interpolation of LSF trajectories, (c) unit fusion, (d) natural speech. The arrows indicate concatenation points in sonorant speech segments.

objective distance between concatenated and natural utterances by 5% (± 1%, $\alpha = 0.05$), while unit fusion reduces the objective distance by 12% (± 1%, $\alpha = 0.05$). This shows that the fused utterances were spectrally closer to natural speech than the linearly smoothed utterances.

The waveforms used in the perceptual and objective test can be accessed at `http://cslu.cse.ogi.edu/tts/demos/ieee_sap01`. In Figure 3.3, spectrograms are shown for one of the sentences, synthesized using the different methods. It is seen that the unit fusion technique removes concatenation mismatch and imposes natural transitions at the concatenation points in sonorant regions. In comparison, the linear interpolation technique imposes smooth but sometimes artificial transitions. For example, the second formant is missing in the transition region in /iʊ/.

3.3 Control of Degree of Articulation

Synthetic speech is often perceived as being over-articulated or "robot-like". This can be attributed to the fact that the generated utterances consist of carefully articulated segments spoken in a discourse-neutral context. On the other hand, utterances generated from large databases of fluent speech often have an uneven quality due to the concatenation of units pronounced with different articulatory effort. In this section, we argue that the quality of synthetic speech can be significantly improved by modifying the spectral dynamics of concatenated units, using the cost function introduced previously. We present results showing that the spectral dynamics of accelerated speech can be successfully controlled to correspond with a more relaxed speaking style.

3.3.1 Background

Studies on vowel reduction and coarticulation show that phonemes in natural speech do not reach the same acoustic targets in all circumstances (Lindblom, 1963; Gay, 1968; Fourakis, 1991; van Bergem, 1993). In formant synthesis, researchers have specified rules to mimic vowel reduction, for example by manipulating formant targets or transition slopes (Klatt, 1987; Kohler, 1990; Granström, 1992). Unfortunately, these rules are not easily applied in the concatenative synthesis framework because they require tracking of formant trajectories in natural speech and control over individual formant frequencies and bandwidths during synthesis.

A model of vowel reduction based on articulatory motor theory was proposed by Lindblom (1990). According to this model, the spectral rate-of-change of speech is governed by the speaker's *articulation effort.* Increased effort leads to faster movements of the articulators, and hence to faster changes in the speech acoustics. Phonetic reduction then can be viewed as the result of an interaction between articulation effort and speaking rate. If the articulation effort is high, target spectral values can be reached even in short phonemes (hyper-articulated speech). If the articulation effort is low, the target spectra may not be reached even in relatively long vowels (hypo-articulated speech).

The Hyper & Hypo model, as it was denoted by Lindblom, motivates a strategy for control of the perceived degree of articulation in synthetic speech. Namely, phonetic reduction can be achieved by constraining the rate-of-change of speech spectra while keeping phoneme durations constant. This can be accomplished using a cost function as in Equation (3.3). Similarly, a hyper-articulated speaking style can be produced by increasing the slopes of spectral transitions, thus enforcing faster movement between phonetic targets.

In the next section, we describe a preliminary experiment to modify the speaking style of synthetic utterances. More sophisticated modifications related to the linguistic structure of a message, such as syllabic stress or word emphasis, are the topic of Chapters 4 and 5.

3.3.2 Reduction of Degree of Articulation in Accelerated Speech

The speaking rate of recorded speech can be increased using PSOLA by removing pitch periods throughout the utterance. Usually, utterances are compressed linearly, i.e., every nth period in each phoneme is dropped to scale the duration by a factor $(n-1)/n$. As a result, approximately the same phonetic targets are reached in accelerated utterances as in the original recordings. According to the Hyper & Hypo theory, however, linear compression of speech units corresponds with the effect of increased articulation effort in natural speech. This is because identical phonetic targets are reached in a shorter time-frame. As a result, accelerated speech often creates the impression of *over*-articulated speech, and becomes unnatural for high speaking rates, i.e., extraordinary effort would be required from a human speaker to produce such acoustic patterns. We therefore propose that a more natural speaking style can be achieved by imposing spectral dynamics consistent with a reduced degree of articulation.

To explore this, we recorded a speaker reading a set of three- or four-digit numbers twice. The first reading was slow and hyper-articulated. The second reading was fast and more relaxed, i.e., hypo-articulated. We then imposed the durations and pitch of the fast readings on the slow readings, thus producing *accelerated* versions of the slow speech. First, the phonemes were time-compressed using PSOLA. Then, dynamic constraints were applied to the speech spectra in order to reduce the perceived degree of articulation.

The dynamic constraints were derived from the smoothed LSF trajectories of the original slow speech. For each pitch-synchronous frame in the accelerated utterance, the derivatives $\Delta^i f_{i,j}$ were extracted from the corresponding point in the slow utterance. These dynamics were imposed on the LSF trajectories of the accelerated utterance by minimizing the cost function in Equation (3.3). This procedure corresponds to maintaining *equal effort* in the accelerated utterances as was used by the speaker in the original slow utterances. To make the accelerated speech sound even more relaxed, it is also possible to scale $\Delta^i f_{i,j}$ in (3.3) by a factor $\beta < 1$. In our experiment, a constant factor $\beta = 0.85$ was applied to the imposed time-derivatives across all sonorant regions of the utterance. In the experiments described in Chapter 5, however, β will be varied as a function of the linguistic structure of a sentence.

3.3.3 Experimental Results

Nineteen utterances were synthesized once using linear time-compression and once using spectral constraints in addition to time-compression. Twenty-two listeners were asked to compare the perceptual quality of the resulting utterances. The same listeners and test setup were used as in Section 3.2. The average perceptual score (CMOS) was 0.28 in favor of the dynamic constraint method ($p < 0.001$, two-tailed t-test), while the highest improvement for a particular sentence was 0.95. We hypothesize that the resulting speech

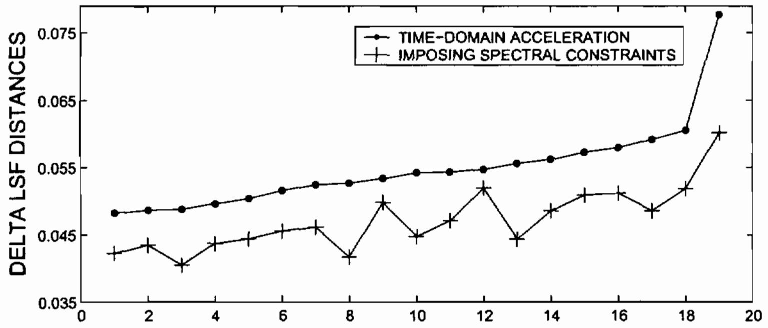

Figure 3.4: Average spectral distances between accelerated speech and natural targets. The distances are reduced with 14% (see text) when dynamic constraints are imposed.

was preferred by human listeners because the perceived degree of articulation was successfully reduced. However, some challenges remain to predict appropriate dynamics for individual phonemes, and to develop an experimental context that more clearly demonstrates the effects of controlling the degree of articulation in synthetic speech.

Another way to test the validity of our approach is to compare the accelerated speech with the natural examples of fast speech, using an objective distance measure. We computed the root-mean-square Euclidean distance between the derivatives of LSF features, extracted from the accelerated utterances and the natural fast speech. The features were obtained from the sonorant speech segments only, and were computed pitch-synchronously using a mel-warped FFT power spectrum. The results are shown in Figure 3.4. It can be seen that the dynamically constrained utterances have a spectral rate-of-change closer to natural speech than the linearly compressed utterances. A regression analysis shows that the distance is reduced by 14% (± 2%, $\alpha = 0.05$) using the spectral modifications. Similarly, for a measure based on delta cepstra, the distances were reduced by 11% (± 2%, $\alpha = 0.05$).

The utterances used for the perceptual test and for the objective measures can be downloaded from `http://cslu.cse.ogi.edu/tts/demos/ieee_sap01`.

3.4 Conclusion

We have presented methods and motivation for control of spectral dynamics in concatenative synthesis. Unit fusion was introduced as a method to reduce spectral mismatch between concatenated units, by imposing spectral constraints derived from independently selected acoustic units. Then, a technique was proposed to control the perceived degree of articulation in synthetic speech, based on the cost function introduced in the unit fusion method. Control of degree of articulation was motivated by a theoretical model which relates the articulation effort to the rate-of-change of speech acoustics.

In the next chapter, the relationship between the linguistic structure of an utterance and the spectral dynamics is investigated further. Our goal is to control the linguistic prominence of syllables and words by modeling the spectral dynamics in conjunction with fundamental frequency and timing, so that increasingly natural and expressive synthetic speech can be generated.

Chapter 4

Effects of prosodic factors on spectral dynamics: Analysis [1]

The effects of prosodic factors on the spectral rate-of-change of vowels are investigated. Thirty two-syllable English words were placed in carrier phrases and read by a single speaker. Liquid-vowel, diphthong, and vowel-liquid transitions were extracted from twenty-four prosodic contexts, corresponding to different levels of stress, pitch accent, word position, and speaking style. The spectral rate-of-change in these transitions was measured by fitting linear regression lines to the first three formants and computing the root mean square of the slopes. Analysis showed that the spectral rate-of-change increased with linguistic prominence, i.e., in stressed syllables, accented words, sentence-medial words, and in hyper-articulated speaking styles. These results support a contextual view of vowel reduction, where the extent of reduction depends both on the spectral rate-of-change and on vowel duration. A numerical model of spectral rate-of-change is proposed, which can be integrated in a system for concatenative speech synthesis, as discussed in Chapter 5.

4.1 Introduction

Prosody refers to supra-segmental aspects of speech such as phrasing, accent placement, stress, and speaking style. Although a number of studies have explored the effects of

[1] **This chapter was submitted as Johan Wouters and Michael W. Macon: "Control of spectral dynamics in concatenative speech synthesis," to the *Journal of the Acoustical Society of America*. This work was supported by the National Science Foundation under Grant No. LIS-0082718 to Michael Macon and Jan van Santen, and by Intel Corporation.**

prosody on the degree of articulation of vowels or *vowel quality* (e.g., Lindblom, 1963; Fourakis, 1991; van Bergem, 1993; Sluijter and van Heuven, 1996), such acoustic-prosodic effects are rarely integrated in automatic speech processing systems. As a result, important cues for robust speech recognition may be ignored, while synthetic speech often sounds over-articulated and requires a higher listening effort than natural speech (van Santen, 1997; Delogu *et al.*, 1998).

In this chapter, the effects of prosodic context on the spectral rate-of-change of vowel transitions are investigated. Nelson (1983) and Moon and Lindblom (1994) showed that the speed of articulatory movements, and hence the spectral rate-of-change, depends on a speaker's *articulation effort*. We believe that the articulation effort in turn depends on prosodic factors, such as the speaking style and the linguistic structure of a message. This motivates our hypothesis that the spectral rate-of-change of phonetic transitions increases with *linguistic prominence*, i.e., with the relative importance of a transition in an utterance, given the linguistic-prosodic context.

In Section 4.2.5, the spectral rate-of-change is defined as the root mean square of the first three formant slopes. We design and analyze a balanced speech corpus, which shows that the spectral rate-of-change increases in stressed *vs.* unstressed syllables, in accented *vs.* unaccented words, in sentence-medial *vs.* sentence-final words, and in hyper-articulated *vs.* hypo-articulated speech, thus providing evidence that the articulation effort increases in linguistically prominent syllables. The results also support a contextual view of vowel reduction, which can be integrated in a concatenative synthesis system. This is the topic of Chapter 5 (also Wouters and Macon, 2001b).

Several researchers have investigated the effects of speaking rate and linguistic stress on the spectral structure of vowels. Lindblom (1963) measured formant frequencies at vowel nuclei in short and long vowels, and defined the difference between the measured values and the formant frequencies reached in long vowels as *target undershoot*. The degree of undershoot could be predicted based on the vowel duration and the distance between the target and the formant value at the vowel onset. Gay (1978), on the other hand, found no effects of speaking rate or stress on vowel quality, even when the segmental durations changed substantially. Van Son and Pols compared formant targets (van Son and Pols,

1990) and formant movements (van Son and Pols, 1992) of Dutch vowels in normal and fast speech. The results contradicted the target undershoot model because the speaker achieved the same formant targets at both speaking rates, possibly by adapting his articulation style (van Son and Pols, 1992).

The target undershoot model regards vowel reduction as a consequence of contextual assimilation (Lindblom, 1963). A competing view on vowel reduction is that vowels shift towards a neutral target in faster speech and in unstressed syllables. This *centralization theory* was investigated by Fourakis (1991), who computed the distance between a neutral vowel target and vowels produced in different conditions of stress and speaking rate. The results did not support the centralization theory in the sense that individual vowels did not come closer to the neutral point in unstressed or fast speech. However, the vowel space shifted and decreased in size for the unstressed and fast conditions. In a recent study, Wrede *et al.* (2000) analyzed the spectral properties of vowels in a large corpus of spontaneous speech. They confirmed that the vowel space became smaller for shorter vowels, while the spectral rate-of-change in vowel onglides and offglides was not affected by the segmental durations.

Van Bergem (1993) reviewed the assimilation vs. centralization argument and concluded that vowel reduction is the result of contextual assimilation, which frequently, but not necessarily, leads to a centralization of formant patterns. Van Bergem also demonstrated that several prosodic factors, apart from speaking rate, produce vowel reduction, such as syllabic stress, sentence accent, and word class. He argued that vowel reduction is not a result of duration changes, as assumed in Lindblom (1963)'s target undershoot model, but duration and vowel quality are both affected by prosodic factors and speaking style.

Strange (1989b,a) investigated the role of spectral dynamics in speech *perception*. She conducted intelligibility experiments with vowels in which the stationary regions or the transition regions were replaced by silence. Both vowel types yielded similar identification rates, and were slightly less intelligible than unmodified vowels. Furui (1986), in a series of experiments, showed that speech segments of 10 ms, excised at the point of maximum spectral change, contain the most important information for identification of consonant-vowel transitions. In spite of the perceptual importance of vowel dynamics, Watson and Harrington (1999) and Pitermann (2000) reported no improvements in automatic vowel classification when dynamic parameters were added to static features. Watson and Harrington (1999) argued that dynamic information, such as spectral rate-of-change, may be related to prosodic aspects of vowels rather than carrying phonetic information.

Moon and Lindblom (1994) reviewed the target undershoot model to account for differences between speaking styles. They modeled the speech motor mechanism as a second order system, and represented the articulation effort by the system input force. The model expressed a linear relationship between the articulation effort and the speed of articulatory movements in the vocal tract. To reflect variations in the articulation effort, Moon and Lindblom added a formant rate-of-change parameter to the undershoot model. In a study of tongue movements in normal and fast speech, Flege (1988) also found that the peak velocity of the tongue was important to predict the undershoot of tongue movements. Similarly, Pitermann (2000) showed that in a database of [iVi] stimuli the formant rate-of-change increased with speaking rate, which reduced the degree of undershoot. In these and other studies (Nelson, 1983; Schulman, 1988; van Bergem, 1993; Watson and Harrington, 1999), changes in tongue velocity or formant rate-of-change are viewed as the result of variations in the articulation effort.

Lindblom (1990)'s Hyper & Hypo Theory views phonetic variation as the result of a trade-off between economy of articulatory movement and communicative requirements. Depending on a speaker's articulation effort, different speaking styles can be produced ranging from sloppy, *hypo*-articulated speech to clear, *hyper*-articulated speech. This presents a new framework for studying vowel reduction. In long vowels or in hyper-articulated speech, the articulators may reach a stationary position corresponding to the vowel's phonetic target. In shorter vowels or in relaxed speech, the time available for the phonetic transition may be insufficient to reach the target, and undershoot occurs. If, for a given vowel duration, speaking effort is increased to speed up articulatory movements, then phonetic targets may be reached even in a short vowel.

In this work, we investigate the hypothesis that a speaker's degree of articulation, as reflected by the rate-of-change of vowel transitions, varies with prosodic structure. Based on the literature review, we expect a positive correlation between linguistic prominence and the spectral rate-of-change. The present work extends Lindblom (1990)'s Hyper & Hypo Theory, in that variations in speaking style are expected *throughout* a spoken utterance as well as across different utterances. A similar view was expressed by de Jong (1995), who measured tongue, jaw, and lip movements in monosyllabic words under varying conditions of stress, and concluded that the data were best explained by considering linguistic prominence as *localized hyper-articulation.*

The present study is limited to prosodic effects on the spectral *time-dynamics* of sonorant speech. Other acoustic correlates of prosody reported in the literature include spectral balance (Sluijter and van Heuven, 1996) and intensity (Waibel, 1986). However, such effects may be more related to the vocal excitation or to a global shift in articulatory movements (e.g., lowered jaw (de Jong, 1995; van Son and Pols, 1992; Traunmüller and Eriksson, 2000)), and are not investigated here. Prosodic effects on non-sonorants, as studied by (van Son and Pols, 1999), are also outside the scope of this study.

4.2 Methods

Three types of vowel transitions were studied: (1) liquid-vowel transitions, (2) diphthong transitions, and (3) vowel-liquid transitions. These transitions correspond to fairly large formant transitions, and were expected to vary appreciably due to changes in articulation effort. Similar transitions were studied by Moon and Lindblom (1994).

The transitions were placed in different prosodic contexts using carrier phrases, according to a balanced experimental design. This text corpus was read by a single speaker. Therefore, the present work aims to draw an accurate picture of the acoustic-prosodic characteristics of *one* speaker. The synthesis experiments in Chapter 5 are based on the same speaker. Investigation of speaker-dependent effects is left for future work.

4.2.1 Prosodic Factors

Previous studies have identified several prosodic factors that affect the degree of articulation of vowels. These include stress, pitch accent, word length, word position, speaking rate, and word class. In this chapter, we investigate the acoustic effects of four factors: (1) syllabic stress, (2) pitch accent, (3) word position, and (4) speaking style. These factors cover a broad range of the phenomena described in the literature, and are a first step in integrating knowledge about the spectral dynamics of vowels in a concatenative speech synthesizer.

Stress is a lexical property of syllables in English words, indicating which syllable is, in some sense, stronger than the others (Sluijter and van Heuven, 1996). Two levels of stress are considered in this work, i.e., syllables receive either primary stress or are unstressed. In previous studies concerning prosodic-acoustic effects, syllabic stress was sometimes confounded with pitch accent. However, syllabic stress and pitch accent may have different acoustic correlates in fluent speech (van Bergem, 1993; Sluijter and van Heuven, 1996). Therefore, stress and accent are investigated independently here.

Several researchers have investigated the acoustic effects of pitch accent, whether or not in conjunction with stress (Lindblom, 1963; de Jong, 1995; Sluijter and van Heuven, 1996; van Bergem, 1993). For the present study, sentences were constructed to contain a single nuclear pitch accent. Words in a sentence are therefore considered either accented or not. Given the fixed structure of the sentences in the database, the non-accented words of interest may precede the accent nucleus by two mono-syllabic words, or follow the accent nucleus by one two-syllable word.

The final phrase boundary of a sentence affects the articulation effort of the word(s) preceding it. Lindblom (1963) varied the acoustic realization of vowels by placing words in a sentence-initial or sentence-final position, while Moon and Lindblom (1994) increased the number of syllables in a word to obtain shortening effects. In the present study, words are considered either sentence-final or sentence-medial. The sentence-medial words of interest are separated from the final sentence boundary by three words.

The effect of speaking rate on vowel quality has been investigated by Flege (1988);

Table 4.1: Prosodic factors

factor	stress	accent	position	style
level 1	primary	accented	medial	clear
level 2	unstressed	non-accented	final	fast
level 3				relaxed

Fourakis (1991); Gay (1978); Pitermann (2000); van Son and Pols (1992) and others. Conflicting results were often found, due to the variety in ***articulation styles*** adopted by the speakers. In some studies, speakers were requested to increase their speaking rate while maintaining accurate pronunciations (van Son and Pols, 1992), or to speak at a tempo indicated by a metronome (Pitermann, 2000). In other cases, speakers spoke at a self-selected fast rate (Gay, 1978; Fourakis, 1991; Flege, 1988). In the present work, we follow Lindblom (1990)'s proposal that speaking styles vary along a continuum from hypo- to hyper-articulated speech. Moreover, we expect that articulation effort and vowel duration behave in a rather independent way, allowing a speaker to articulate in a fast and clear manner, or in a slow and sloppy manner, or any combination in between. To investigate this, the speaker in our experiment read the same set of sentences in three different speaking styles, i.e., clear, fast, and relaxed. The recording protocol is explained in more detail in Section 4.2.3.

The prosodic factors and their levels are summarized in Table 4.1. To apply a balanced experimental design, $2 \times 2 \times 2 \times 3 = 24$ prosodic conditions need to be considered for each phonetic transition in the database.

4.2.2 Speech Corpus

The speech database was based on 30 two-syllable words with an ambiguous stress pattern in English. The words were selected from a lexicon to include at least one vowel transition involving a diphthong or a liquid. Most of the selected words have primary stress on the first syllable when used as a noun (e.g. abstract), and primary stress on the second syllable when used as a verb (e.g. abstract). The complete list is shown in Table 4.2.

The selected words were placed in meta-linguistic carrier phrases of the form "Please

Table 4.2: Two-syllable words with ambiguous stress pattern

abstract	impulse	research
chauffeur	increase	retake
debut	insert	rewrite
digest	incite/insight	romance
discharge	insult	surcharge
eighteen	monroe	surveys
escort	narrates	transform
exploit	overt	transport
ferment	permit	trustee/trusty
import	pervert	upright

produce ... for him again" and "The speaker will now produce ...". The resulting sentences are grammatically correct, independent of the syntactic properties of the inserted word. The first carrier phrase allows sentence-medial insertions of the target word and the second carrier phrase allows sentence-final insertions.

For each two-syllable word, eight typographically different sentences were generated, corresponding to different conditions of stress, accent, and sentence position. An example is shown in Table 4.3. Stressed syllables are underlined, while accented words appear in capitals. The sentences were repeated in three different speaking styles, producing 24 sentences per target word. Each sentence was recorded three times to improve the estimation of the spectral rate-of-change parameter to be discussed in Section 4.2.5. The total number of utterances in the database is therefore $30 \times 24 \times 3 = 2,160$.

4.2.3 Recording

The corpus was recorded in three sessions, each session corresponding to a particular speaking style. Sentences were grouped in blocks of eight per two-syllable word. These blocks were repeated three times per session and were randomly ordered per repetition. In the first repetition, the sentences within each block appeared as in Table 4.3. In the second repetition, the order of stressed and unstressed syllables was reversed. In the third repetition, the sentence-medial and sentence-final conditions were reversed. This structured ordering per sentence block made it easier for the speaker to vary the desired

Table 4.3: Presentation of the word "abstract" in different prosodic contexts, covering all combinations of stress, accent, and word position levels. Stressed syllables are underlined. Accented words appear in capitals.

1. Please produce ABSTRACT for him again.
2. Please produce abstract for him AGAIN.
3. The speaker will now produce ABSTRACT.
4. The speaker will NOW produce abstract.
5. Please produce ABSTRACT for him again.
6. Please produce abstract for him AGAIN.
7. The speaker will now produce ABSTRACT.
8. The speaker will NOW produce abstract.

prosodic factors for each new sentence.

The sentences were recorded by my research advisor Michael Macon who speaks a northern dialect of American English without any strong regional influence. The speaker was required to correctly assign stress and pitch accents according to the typography of the sentences. This was achieved quite naturally after initial practice, and was helped by presenting the sentences in the order described above, as opposed to a completely random presentation. A sentence was repeated whenever the reading was not satisfactory to the speaker or to the experimenter.

The speaking styles were varied as follows. In the first session, the speaker read the sentences slowly and clearly, corresponding to a hyper-articulated speaking style. In the second session, the speaker read the sentences at a faster, but still comfortable rate. Post-hoc analysis showed a 24% decrease in the vowel durations for the second session compared with the first session. In the third session, the speaker read the sentences in a more relaxed way than either the first or the second session. This resulted in a decrease in vowel durations of 30% compared with the first session.

The utterances were recorded digitally at a sampling frequency of 48 kHz and 16 bits per sample in a sound-insulated booth, using a large-capsule condenser microphone. A laryngograph signal was simultaneously recorded to allow accurate marking of the pitch onset times of the speech for pitch-synchronous spectral analysis. After the recording session, the speech and laryngograph signals were downsampled to 16 kHz for further processing.

4.2.4 Transition regions

The transitions of interest for this study were located in the sonorant regions of the two-syllable words. The following three *transition types* were considered:

liquid-vowel transitions; for example /ɹi/ in "**re**search"

diphthong transitions; for example /aɪ/ in "d**i**gest"

vowel-liquid transitions; for example /ʌl/ in "ins**ul**t"

For the two-syllable words in Table 4.2, most liquid-vowel or vowel-liquid transitions involve /ɹ/, while transitions with /l/ also occur. The diphthong transitions are mostly towards /i/, although the /oʊ/ and /ju/ diphthongs are also present. The transitions are summarized in Table 4.4. Some two-syllable words were included in Table 4.2 to cover /ɝ/ transitions, as in "ins**er**t". However, since the realization of /ɝ/ corresponds more to a monophthong than to a diphthong, /ɝ/ transitions are not included in Table 4.4 or in the analyses presented below. Hence, thirty-nine phonetic transitions extracted from different two-syllable words remain.

The liquid-vowel transitions occur in syllable onsets, while the vowel-liquid transitions occur in syllable coda's. In the remainder of this study, the three transition types will therefore be referred to as *onset*, *diphthong*, and *coda* transitions.

All transitions were excised from the recorded sentences using an automatic segmentation algorithm recently developed by Hosom (2000). This algorithm provides an objective criterion to partition the vowels into stationary and transition regions, thus avoiding possible inconsistencies in segmenting the transitions by hand. The segmentation system has reported a phoneme-level accuracy of 92.5% within 20 ms of manual labels for TIMIT sentences, which is close to agreements between human labelers reported in the literature (see Hosom (2000)). Inspection of the phoneme labels for the present corpus revealed no obvious segmentation errors.

The automatic segmenter aligns Hidden Markov Model (HMM) states with the speech waveform, using a phonetic transcription of the utterance. Monophthongs are represented

Table 4.4: Overview of the phonetic transitions extracted from different two-syllable words.

liquid-vowel		diphthong		vowel-liquid	
/l/-vowel		vowel-/i/		vowel-/l/	
/lɔɪ/	1	/eɪ/	5	/ʌl/	2
/ɹ/-vowel		/aɪ/	4	vowel-/ɹ/	
/ɹi/	4	/ɔɪ/	1	/ɔɹ/	4
/ɹæ/	3	vowel-/u/		/aɹ/	2
/ɹoʊ/	2	/oʊ/	4	/ɛɹ/	1
/ɹaɪ/	2	/ju/	1	/iɹ/	1
/ɹeɪ/	1				
/ɹʌ/	1				
	14		15		10

by three HMM states and liquids by two states. Diphthongs are modeled as a sequence of two monophthongs. Therefore, we defined the starting point b of each transition region *p1-p2* as the beginning of the last HMM state representing phoneme *p1*, and the end point e as the end of the first HMM state representing phoneme *p2*. To prevent estimation of the spectral rate-of-change in very short speech segments, b and e were corrected to be at least 20 ms away from the phoneme boundary between *p1* and *p2*. In the next section, a spectral rate-of-change measure is developed which reflects the spectral dynamics at the center of a transition region, while being relatively insensitive to variations in the region boundaries.

4.2.5 Spectral Rate-of-Change

To investigate the effects of the prosodic factors, a measure of the spectral rate-of-change must be obtained for the transition regions described above. We denote this measure as *RoC*. One possible measure would be to extract spectral features such as mel-frequency cepstrum coefficients (MFCC) at equidistant frames and compute the average Euclidean distance between successive frames across each transition region. However, preliminary experiments showed that such estimates were quite noisy and did not correspond well with visual estimations of the rate-of-change in spectrograms. This is due partly to the fact that the MFCC distance reflects all spectral changes, including variations in spectral balance and frame-to-frame jitter.

For this study, a measure was needed that reflects the articulatory movements over the course of a phonetic transition. Therefore, we adopted a spectral rate-of-change measure based on formant trajectories, which follow the resonant frequencies associated with successive vocal tract configurations. Formant tracking is a known hard problem in speech processing because the resonant frequencies of the vocal tract are not always observable in the output speech signal (Broad and Clermont, 1989; Laprie and Berger, 1996; Welling and Ney, 1998; Acero, 1999). However, for the present task the formants needed to be identified only in short regions of sonorant speech. A simple algorithm which finds the best path through a set of LPC poles was therefore able to correctly track the desired formants for most of the data. The LPC poles were computed from a mel-warped power spectrum, obtained via pitch-synchronous sinusoidal analysis of the speech signal (Wouters and Macon, 2000). About 40% of the transitions were manually checked, revealing tracking errors in less than 1.5% of the transitions.

After formant tracking, linear regression lines were fit to the formant trajectories in each transition region. A bell-shaped weighting function was applied to improve the fit towards the center of the transition regions, making the regression lines less sensitive to the placement of the transition boundaries. Since the transition regions were rather short (average 59 ms), the regression lines closely approximated the formant trajectories in most cases. The slopes α_i of the regression lines describe the formant movements at the transition point between two phonemes, and can be assumed proportional to the peak velocity of the articulatory movement between the phoneme targets. Therefore, the regression line slopes may be considered as a measure of the articulation effort expended at a particular phoneme transition (Nelson, 1983).

Finally, the spectral rate-of-change of a transition region was defined as the normalized p-norm of the absolute slopes $|\alpha_i|$ for a set of formants S_N:

$$RoC = \left(\frac{1}{N} \sum_{i \in S_N} |\alpha_i|^p \right)^{1/p} . \tag{4.1}$$

Different formant combinations were investigated using the analysis of variance (ANOVA) model to be discussed in Section 4.3.2. While the ANOVA results were similar, the main

Table 4.5: Percentage of variance explained (R^2) by the ANOVA model for different definitions of the spectral rate-of-change.

	RoC	$R^2 \times 100\%$
N=1	F1	51.08
	F2	70.86
	F3	56.73
N=2, p=2	F1,F2	70.54
	F2,F3	75.56
	F1,F3	66.30
N=3, p=2	F1,F2,F3	**75.68**
N=3, p=1	F1,F2,F3	74.68
	$\sqrt[3]{\lvert F1F2F3\rvert}$	56.36

effects of the prosodic factors became more significant when more formant slopes were included in the measure. The percentage of variance explained by ANOVA for different definitions of *RoC* also increased with the number of formants N, as shown in Table 4.5. One explanation why the model is more powerful for higher N is that the prosodic factors determine variations in the *articulatory movements*, which are reflected in different formants depending on the articulator(s) involved (e.g. effect of retroflex movement on F3).

For $N = 3$, the $p = 1$ norm yielded less significant main effects and modeling power than the quadratic norm, as did a multiplicative combination of the three formant slopes (see Table 4.5). Therefore, in the remainder of this chapter, analysis will be based on the root mean square of the first three formant slopes, i.e., $p = 2$ and $N = 3$ in Equation 4.1. Since the slopes were estimated from a mel-warped spectrum, the transitions of lower formants are weighted more heavily, in correspondence with the frequency resolution of human hearing. No additional weighting of the formant slopes was performed.

4.3 Statistical Analysis

In this section the effects of the factors stress, accent, word position, and speaking style on the spectral rate-of-change of transitions in the database are investigated. First, the main

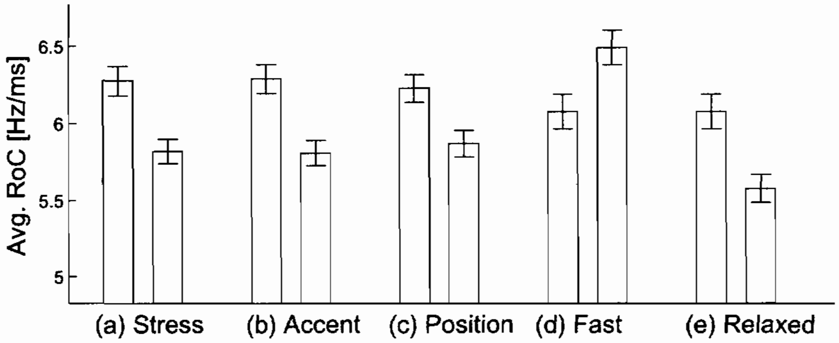

Figure 4.1: Average spectral rate-of-change [Hz/ms] for different subsets of the database: (a) stressed vs. unstressed, (b) accented vs. unaccented, (c) sentence-medial vs. sentence-final, (d) clear vs. fast style, (e) clear vs. relaxed style.

effects of the prosodic factors are studied by comparing *RoC* means computed over specific subsets of the data. This provides a first insight in the data and suggests a refinement of the prosodic factor word position. Then, analysis of variance is performed to study the effects of interactions between the prosodic factors. Finally, a numerical model of the prosodic effects is proposed.

4.3.1 Main Effects

In Figure 4.1 the main effects of the prosodic factors are investigated by comparing the mean *RoC*'s computed for each level of a prosodic factor. The estimated standard deviations of the means are also shown. The differences are significant in all cases ($p < 0.05$, double-sided t-test), indicating that *RoC* is *higher* on average in stressed vs. unstressed syllables, in pitch-accented vs. unaccented words, and in sentence-medial vs. sentence-final words. For the three speaking styles adopted in the corpus, *RoC* *decreases* in the relaxed style compared with the clear speaking style, and *increases* in the fast speaking style.

Figure 4.2 shows the effect of the sentence-medial vs. sentence-final word position for onset, diphthong, and coda transitions, respectively in the first and second syllable of the target words. The differences are significant for the transitions in the second syllable ($p < .05$, single-sided t-test). Hence, the spectral rate-of-change decreases in sentence-final

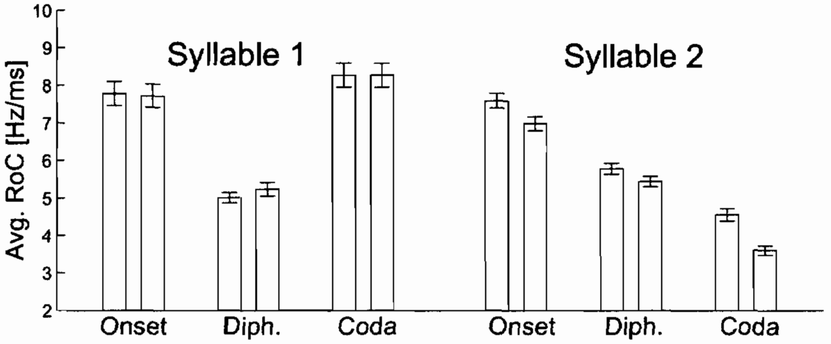

Figure 4.2: Spectral rate-of-change for sentence-medial vs. sentence-final transitions, from left to right: Syllable 1: onset, diphthong, coda; Sylable 2: onset, diphthong, coda. The effect of the phrase boundary is significant only ($p < .05$, single-sided t-test) for the transitions in the second, sentence-final, syllable.

transitions, but this effect is limited to the final syllable of a sentence, and is strongest for the syllable coda transition at the end of a sentence. For further analysis, we redefine the factor word position so that transitions in the second syllable of target words can be sentence-final, but transitions in the first syllable are always considered sentence-medial.

4.3.2 Analysis of Variance

To study both the main effects and the interactions between the prosodic factors, analysis of variance (ANOVA) was performed. A six-way ANOVA was computed with stress, accent, position, style, transition type, and transition identity as independent variables and *RoC* as the dependent variable. Transition type (onset, diphthong, or coda) was added as an independent variable after exploratory analysis showed the importance of this factor. Hence, there are two levels for stress, accent, and position, three levels for style and transition type, and 39 levels for transition identity (see Table 4.4). The *RoC* measurements are balanced for the independent variables except for transition type and position. The latter is due to constraining the sentence-final word position level to transitions in the second syllable.

The ANOVA model describes the dependent variable as a sum of terms. However, in the case of spectral rate-of-change, a multiplicative model of the prosodic factors may be

Table 4.6: Analysis of variance (ANOVA) of the logarithm of the rate-of-change measurements. The independent variables are transition identity, stress, accent, word position, speaking style, and transition type. For each factor, the degrees of freedom, sum of squares, F-value, and significance are shown. Significant effects (< .05) are marked with (*).

	Df	SS	F	Pr(F)
Main effects				
identity	38	724.7	206.4	0.000*
stress	1	2.4	26.4	0.000*
accent	1	3.0	33.1	0.000*
position	1	7.3	80.0	0.000*
style	2	8.6	46.6	0.000*
2-way interactions				
trans:stress	2	3.8	20.5	0.000*
trans:accent	2	1.6	9.0	0.000*
trans:position	2	2.9	15.9	0.000*
trans:style	4	16.9	45.7	0.000*
stress:accent	1	1.0	11.6	0.000*
stress:position	1	0.0	0.4	0.500
stress:style	2	0.2	1.5	0.209
accent:position	1	0.0	0.2	0.648
accent:style	2	0.0	0.4	0.656
position:style	2	1.3	7.2	0.000*
3-way interactions				
trans:stress:accent	2	0.0	0.0	0.947
trans:stress:position	2	0.3	1.8	0.159
trans:stress:style	4	0.1	0.4	0.799
trans:accent:position	2	0.0	0.0	0.970
trans:accent:style	4	0.4	1.1	0.320
trans:position:style	4	0.5	1.4	0.207
stress:accent:position	1	0.1	1.1	0.285
stress:accent:style	2	0.1	0.7	0.453
stress:position:style	2	0.2	1.3	0.252
accent:position:style	2	0.1	0.8	0.445
4-way interactions				
trans:stress:accent:position	2	0.2	1.5	0.203
trans:stress:accent:style	4	0.2	0.6	0.627
trans:stress:position:style	4	0.1	0.3	0.834
trans:accent:position:style	4	0.1	0.2	0.887
stress:accent:position:style	2	0.0	0.4	0.647
5-way interactions				
trans:stress:accent:position:style	4	0.0	0.1	0.957
Residuals	2700	249.7		

more appropriate. Such a model assumes that rapid articulatory movements are attenuated more as a result of prosodic factors, while slow movements vary less. To illustrate, if a spectral transition of 80 Hz/s decreases to 60 Hz/s for a certain prosodic factor, then with a multiplicative model a transition of 8 Hz/s decreases to 6 Hz/s. In the case of an additive model, a decrease in *RoC* by 20 Hz/s, regardless of the original *RoC* value, could lead to counter-intuitive results: a transition of 8 Hz/s would be inverted to -12 Hz/s, unless additional constraints are imposed.

We computed the ANOVA both for *RoC* and for log(*RoC*). The results were comparable, but the interactions between the prosodic factors were somewhat less significant for log(*RoC*), indicating that the logarithmic transformation better separates the main effects of the prosodic factors. Therefore, in Table 4.6 the results for log(*RoC*) are reported. The same transformation was used for the results in Table 4.5. In the remainder of this section, marginal means will be reported in percent change of *RoC*, corresponding to linear changes of log(*RoC*).

The analysis of variance confirms the main effects of stress, accent, position, and style shown in Figure 4.1. The marginal means of the main effects indicate that *RoC* increases by 5.9% in stressed vs. unstressed syllables, by 6.6% in accented vs. non-accented words, by 9.5% in sentence-medial vs. sentence-final transitions, by 4.7% in fast vs. clear speech, and by 8.7% in clear vs. relaxed speech.

There are significant two-way interactions between the transition type and each of the prosodic factors. These interactions are now discussed in detail. A graphical representation of the prosodic effects for different transition types is given in Figure 4.3, although the numerical results of the ANOVA are somewhat different from the quantitative model discussed there.

The interaction between transition type and stress increases the main effect of stress for onset transitions, while neutralizing the effect for diphthong and coda transitions. When the marginal means of the interaction are combined with the main effect, *RoC* increases by 16% for stressed vs. unstressed conditions in onset transitions. Diphthong transitions decrease by 2%, and coda transitions increase by 4%.

A similar interaction is observed between transition type and accent. For onset transitions, the main effect of pitch accent is reinforced (+13%), while it is smaller for diphthong (+3%) and coda (+5%) transitions.

The interaction between transition type and word position is in correspondence with Figure 4.2 earlier. The effect of word position becomes smaller in onset (+7%) and diphthong (+4%) transitions, and larger in coda (+20%) transitions.

The interaction between transition type and speaking style results in slower onset transitions in the fast speaking style compared to the clear speaking style (-9%), and in dramatically slower onset transitions in the relaxed speaking style compared to the clear speaking style (-34%). Diphthong transitions are more rapid in the fast (16%) and relaxed (10%) styles, while coda transitions are more rapid in the fast style (+9%) and practically the same in the relaxed style (-1%), compared to the clear style. Flege (1988) observed the same asymmetry between vowel onset and coda transition rates. For several speakers, the velocities of onset tongue movements decreased in fast speech compared with normal-rate speech, while the velocities of coda movements increased. We believe this may be due to a hierarchical difference between syllable onsets and coda's, which causes the dynamics of coda transitions to be dominated by preceding and following onset movements.

The interaction between word position and speaking style indicates that the main effect of word position is reinforced in the relaxed speaking style (+19%), while it remains small in the fast speaking style (-3%). The interaction between accent and stress introduces modest (±2%) corrections to the main effects.

All other interactions between prosodic factors, including three-, four-way, and five-way interactions, were not significant. When two-way interactions between transition *identity* and stress, accent, or position were computed, they were found to be highly significant ($p < 0.0001$). However, since the interactions with transition identity have a large number of degrees of freedom, such interactions were excluded from the analysis shown in Table 4.6 to avoid overfitting.

In Table 4.6, the percentage of variance explained by transition identity alone is 70.6%. The percentage of variance explained by transition identity and the four prosodic factors with their interactions is 75.7%, or the prosodic factors account for 17% of the remaining variance. This is explored further in the next section.

4.3.3 A Quantitative Model of Prosodic Effects on Spectral Dynamics

Based on the analysis of variance, we propose a simple multiplicative model to predict the effects of prosodic context on the rate-of-change of vowel transitions in natural speech. Due to its form and to limit the number of parameters to be estimated, the model does not take into account interactions between the prosodic factors, which were shown to be significant for stress and accent, and for word position and speaking style. The effect of the transition type, however, is accounted for by estimating separate parameters for each transition type, i.e., syllable onset, diphthong, and coda transitions.

An advantage of the proposed model compared with the ANOVA model is that the parameters can be estimated without performing a logarithmic transformation. The distribution of the ANOVA prediction errors vs. the fitted values is non-white for log(RoC), showing increasingly large errors for lower RoC values. This may be because slow spectral transitions ($RoC \approx 0$) are not so much affected by the prosodic factors, while the logarithmic transformation boosts the importance of such transitions. To avoid using log(RoC), the parameters in the present model were estimated by a non-linear minimization of the prediction error.

The proposed model describes RoC as the product of five factors:

$$\widehat{RoC} = K_p \times \alpha(S) \times \beta(A) \times \gamma(P) \times \delta(Y) \qquad (4.2)$$

K_p is determined by the transition identity, which reflects the phonetic context of a transition. The factors α to δ have a constant value depending on the level of stress (S), accent (A), word position (P), and speaking style (Y), respectively. These factors were estimated for each transition type independently.

The model in Equation 4.2 requires estimation of one parameter for each level of a prosodic factor. We chose to normalize the prosodic factors by setting the parameters corresponding to the unstressed, unaccented, sentence-final condition in the clear speaking style to 1, i.e., $\alpha(S{=}0) = 1$, $\beta(A{=}0) = 1$, $\gamma(P{=}\text{final}) = 1$, $\delta(Y{=}\text{clear}) = 1$. The remaining parameters were then estimated by minimizing the squared error between the

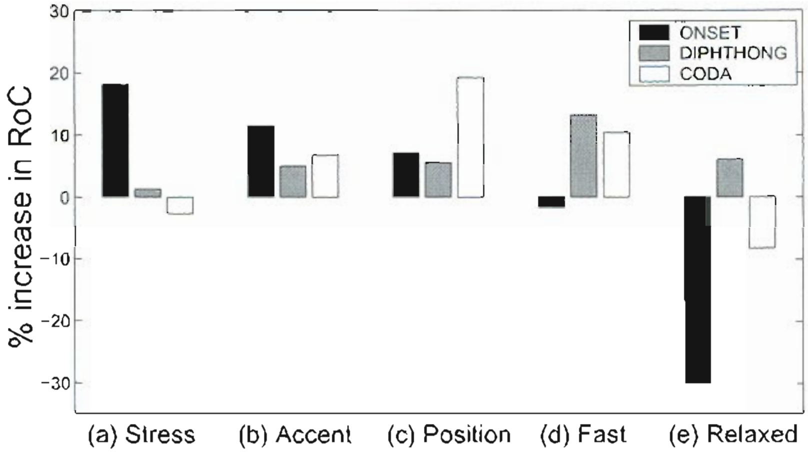

Figure 4.3: Relative increase in spectral rate-of-change (*RoC*) for syllable onset, diphthong, and coda transitions, when comparing (a) stressed vs. unstressed syllables, (b) accented vs. non-accented words, (c) sentence-medial vs. sentence-final transitions, (d) fast vs. clear speech, and (e) relaxed vs. clear speech.

predicted *RoC*'s and the experimentally found values, using a gradient-descent algorithm. This yielded the results shown in Figure 4.3. In this figure, the parameters were normalized to $(\phi - 1) \times 100\%$, $\phi = \alpha..\delta$, to represent the relative increase in *RoC* in stressed vs. unstressed syllables, accented vs. non-accented words, sentence-medial vs. sentence-final transitions, fast vs. clear speech, and relaxed vs. clear speech. The effects are shown for syllable onset, diphthong, and coda transitions, respectively. The numerical values are given in Table 4.7.

Table 4.7: Estimated parameters for Equation 4.2 when $\alpha(S{=}0)$, $\beta(A{=}0)$, $\gamma(P{=}\text{ final})$, and $\delta(Y{=}\text{ clear})$ are normalized to 1. The parameters in this table correspond with the values in Figure 4.3 via $(\phi - 1) \times 100\%$, $\phi = \alpha..\delta$ (see text).

	Onset	Diphthong	Coda
$\alpha(S{=}1)$	1.18	1.01	0.97
$\beta(A{=}1)$	1.11	1.05	1.07
$\gamma(P{=}\text{ medial})$	1.07	1.05	1.19
$\delta(Y{=}\text{ fast})$	0.98	1.13	1.10
$\delta(Y{=}\text{ relaxed})$	0.70	1.06	0.92

The percentage of variance (R^2) explained by the multiplicative model is highest for the syllable onset transitions and for the coda transitions. Using only the transition identities (parameters K_p in Equation 4.2), 65% of the variance is explained for the onset transitions. This increases to 78% when the prosodic factors are brought into the model. Hence, the prosodic factors explain 37% of the remaining variance. For the coda transitions, 65% of the variance is explained by K_p, improving to 70% using the prosodic factors. For the diphthong transitions, the percentage of variance explained by K_p is 51%, increasing to 54% when the prosodic factors are considered.

A considerable part of the variance left unexplained is due to variations between the three repetitions in the corpus. This part of the error can be removed by taking the average of the three *RoC* measurements for each prosodic condition, prior to estimating the model parameters. While the parameters obtained from this estimation are similar to the values in Table 4.7, the percentage of variance explained increases to 88.5% for onsets, 85% for codas, and 70% for diphthongs. The prosodic factors explain 56% of the variance not explained by K_p for onsets, 28% for codas and 10% for diphthongs.

Other sources of variance not modeled by Equation 4.2 are the interactions between the prosodic factors, two of which were shown to be significant using ANOVA, and the interactions between the prosodic factors and the transition identities. Also, some model error may be attributed to inaccuracies in the *RoC* measurements, for example during the segmentation of the transition regions, during formant tracking, or to inconsistencies in the speaker's articulation strategy.

4.4 Discussion

The results of this study suggest that the spectral rate-of-change of vowels increases in stressed syllables, as well as in accented and sentence-medial words, and in clearly articulated speech. As described in the introduction, spectral rate-of-change is related to the articulation effort excercised by a speaker (Nelson, 1983; Moon and Lindblom, 1994; Flege, 1988). Hence, our results support the hypothesis that the degree of articulation of vowels increases with linguistic prominence. This resonates with de Jong (1995)'s view of prominence as *localized hyper-articulation*.

Three types of spectral transitions were investigated: liquid-vowel, diphthong, and vowel-liquid transitions. Lindblom (1963), and others, showed that the degree of formant undershoot depends on the distance between the vowel target and the onset frequency at the consonant-vowel transition. Therefore, large formant transitions were preferred here to investigate the effects of prosodic context on the formant rate-of-change. We expect that variations in articulation effort, driven by changes in the prosodic context, may have similar effects on other vowel transitions. However, our data also showed significant interactions between the phonetic transition identity and the prosodic factors.

Two causes can be identified for the interactions between transition identity and the prosodic factors. Firstly, the relationship between articulatory movements and the spectral rate-of-change depends on the articulators involved. Hence, variations in articulation effort or in articulatory movements may affect certain spectral transitions more than others. Also, the prosodic context may have less impact on "easy" articulatory movements, which are close to a minimum expenditure of articulation effort (Nelson, 1983). Secondly, since the spectral transitions in this study were extracted from different English words, there may be effects of phonetic context outside the investigated diphone transitions. For example, competing demands on one articulator can constrain the spectral rate-of-change of a vowel transition. This could be the case for /oʊ-m/ in "romance", which features lip rounding in /oʊ/ followed by a labial closure. Relatively large *RoC* values were measured for this transition, indicating that the prosodic effects may be dominated by the labial closure.

The prosodic effects on the spectral rate-of-change are rather small in several cases (see Figure 4.3). Hence, in a first approximation the spectral rate-of-change of vowels might be considered invariant across prosodic contexts and for different vowel durations. This is consistent with several observations in the literature (e.g., Gay, 1978; Hertz, 1991; Wrede *et al.*, 2000; Lindblom, 1963). On closer investigation, however, the spectral rate-of-change increases with the linguistic prominence of a phonetic transition. The effect becomes larger when several prosodic factors combine. For example, a transition in a stressed, accented syllable in clear speech evolves 71% faster than an unstressed, unaccented transition in relaxed speech, using Equation 4.2 and the parameters in Table 4.7.

The present results add support for a *contextual model of articulation.* According to this model, the extent of vowel reduction or the degree of coarticulation between sonorant phonemes depends on (1) the phonetic context, (2) the phoneme durations, and (3) the spectral rate-of-change of the phonetic transitions. Both durations and spectral rate-of-change are determined by the prosodic context, via different mechanisms. While the spectral rate-of-change varies through short-term changes in a speaker's articulation effort, segmental durations can be viewed as the result of changes in the planning of articulatory gestures (Browman and Goldstein, 1992).

The contextual model of articulation is amenable to algorithms for concatenative synthesis. In Chapter 5, methods are discussed to control the spectral rate-of-change of concatenated speech segments, by imposing dynamic constraints on the spectral trajectories. As a result, different degrees of reduction can be achieved for units of a given database. The synthetic speech corresponds to a more natural articulation, which improves the perceptual quality.

Our analysis of the spectral dynamics in natural speech stands in contrast with assumptions of vowel time-invariance commonly made in speech synthesis and recognition systems. In most synthesis algorithms, vowel durations are controlled by uniformly time-stretching or time-compressing recorded speech units, for example using pitch-synchronous overlap-add (Moulines and Charpentier, 1990). As a consequence, the spectral rate-of-change increases in shortened vowels, and decreases in elongated vowels. Since vowel durations increase with stress and accent (van Santen, 1992), and in clear speech, time-invariance imposes a decrease in the spectral rate-of-change in such cases. This is contrary to the behavior observed in the present data.

The difference between uniform time-warping and natural duration changes is illustrated in Figure 4.4 for the present database. Average vowel durations are compared for minimal pairs differing in stress, accent, position, and speaking style (Figure 4.4, Top). The duration differences are consistent with results in the literature (van Santen, 1992). Using uniform time-scale modification to transform linguistically prominent syllables to their less prominent counterparts causes an increase in *RoC* for each prosodic factor except for word position. This can be compared in Figure 4.4 (Bottom) with the actually measured *RoC*

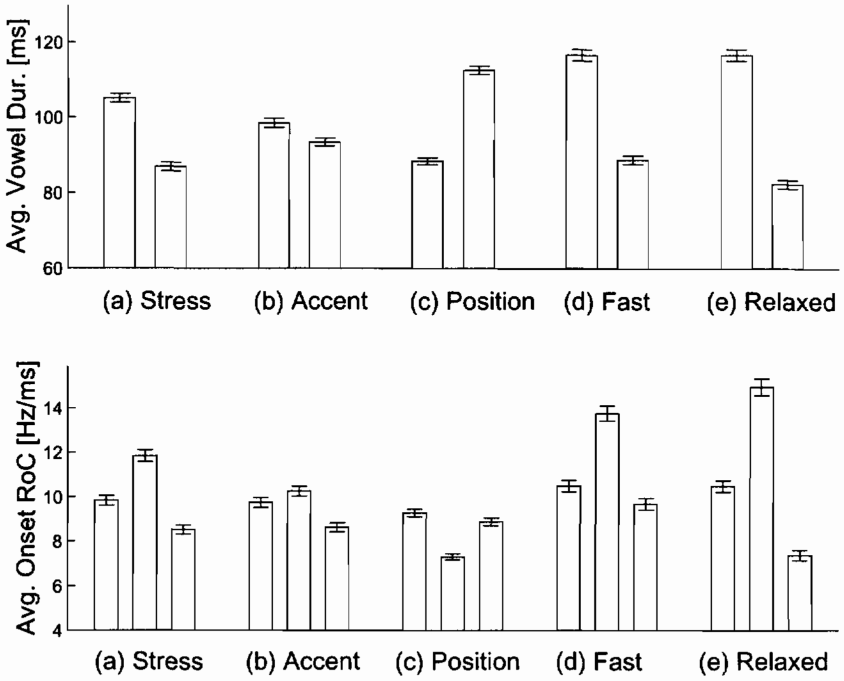

Figure 4.4: Comparison of spectral dynamics in natural speech vs. uniform time-warping. Top: Average vowel durations [ms] for (a) stressed vs. unstressed, (b) accented vs. non-accented, (c) sentence-medial vs. sentence-final, (d) clear vs. fast, and (e) clear vs. relaxed speech conditions. Bottom: Effect of time-warping linguistically prominent syllables using scaling factors derived from top. Average rate-of-change [Hz/ms] of syllable onset transitions are shown for (a) stressed, warped, and unstressed, (b) accented, warped, and non-accented, (c) sentence-medial, warped, and sentence-final, (d) clear, warped, and fast, and (e) clear, warped, and relaxed speech conditions.

values for onset transitions. Except for word position, uniform time-warping increases the spectral rate-of-change where it should in fact decrease. The resulting speech is perceived as over-articulated.

The current results are based on recordings from one speaker. Since different speakers may employ different articulation strategies depending on prosodic context (de Jong, 1995; Flege, 1988; Moon and Lindblom, 1994; Widera, 2000), speaker-dependent effects should also be investigated. However, the present study fits in a synthesis paradigm that attempts to closely reproduce the voice of a specific speaker as opposed to creating a statistical average of speakers. Synthesis experiments reported in Chapter 5 are based on recordings from the same speaker investigated in this study.

4.5 Conclusion

An experimental database was designed to study the effects of prosodic factors on the spectral time-dynamics of glide-vowel, diphthong, and vowel-glide transitions. The results indicated that the spectral rate-of-change increases with linguistic prominence, i.e., in stressed syllables, in accented words, in sentence-medial words, and in hyper-articulated speaking styles. The rate-of-change increased most in glide-vowel transitions, which occur at the syllable onset, while variations in diphthong and vowel-glide transitions were less predictable.

The results in this chapter suggest that, for a given phonetic context, vowel reduction depends on the interaction between the phoneme duration and the spectral rate-of-change. The rate-of-change reflects the articulation effort used by the speaker, while phoneme durations are determined by the timing of discrete articulatory gestures. Both factors depend on the prosodic context. We have developed a numerical model of the effects of prosodic factors on the spectral rate-of-change, which can be integrated in algorithms for automatic speech recognition or synthesis. This is the topic of the next chapter.

Chapter 5

Effects of Prosodic Factors on Spectral Dynamics: Synthesis [1]

In Chapter 4, the effects of prosodic factors on the spectral rate-of-change of phoneme transitions were analyzed for a balanced speech corpus. The results showed that the spectral rate-of-change, defined as the root-mean-square of the first three formant slopes, increased with linguistic prominence, i.e., in stressed syllables, in accented words, in sentence-medial words, and in clearly articulated speech. In this chapter, an initial approach is described to integrate the results of Chapter 4 in a concatenative synthesis framework. The target spectral rate-of-change of acoustic units is predicted based on the prosodic structure of utterances to be synthesized. Then, the spectral shape of the acoustic units is modified according to the predicted spectral rate-of-change. Experiments show that the proposed approach provides control over the degree of articulation of acoustic units, and improves the naturalness and intelligibility of concatenated speech in comparison to standard concatenation methods.

5.1 Introduction

In this chapter, an approach is presented to improve prosodic modification of acoustic units, by controlling the *degree of articulation* of sonorant phonemes. Hence, phenomena

[1] This chapter was submitted as Johan Wouters and Michael W. Macon: "Effects of Prosodic Factors on Spectral Dynamics. Part II: Synthesis," to the *Journal of the Acoustical Society of America*. This work was supported by the National Science Foundation under Grant No. LIS-0082718 to Michael Macon and Jan van Santen, and by Intel Corporation.

such as vowel reduction or co-articulation between vowels, liquids, and glides can be produced at synthesis time. Such control has significant potential to improve the naturalness and intelligibility of concatenated speech, since units selected from small databases often produce over-articulated or "choppy" speech, while units selected from a large database often produce an inconsistent articulation or a "patchwork" quality. In the proposed approach, the degree of articulation of sonorant units is modified depending on the prosodic structure of the target utterance, so that a wide variety of utterances can be synthesized from a given database. The approach is based on the analysis conducted in Chapter 4 (also Wouters and Macon, 2001c).

In formant synthesis, control of the degree of articulation of vowels is achieved relatively easily by re-defining formant targets in certain phonetic and prosodic contexts (Klatt, 1987; Kohler, 1990). However, no consensus has emerged for predicting the desired formant targets depending on the prosodic or phonetic context, and description of the complete formant trajectories presents further challenges (Hertz, 1991; van Bergem, 1993). In concatenative synthesis, reducing formant targets is even more difficult since identification of formant trajectories in recorded speech is not robust, and the trajectories typically cannot be modified by adjusting a small set of system parameters.

In Chapter 4, the spectral dynamics of vowel and liquid transitions were analyzed for different prosodic contexts. The spectral rate-of-change, defined as the root mean square of the first three formant slopes, increased in linguistically prominent phoneme transitions, i.e., in stressed syllables, in accented and sentence-medial words, and in clearly articulated speech. The findings suggested that the speaker's articulation effort increased with linguistic prominence, resulting in faster articulatory movements and hence in a higher spectral rate-of-change (Nelson, 1983; Moon and Lindblom, 1994).

The results in Chapter 4 supported a contextual model of articulation. According to this model, the degree of vowel reduction or co-articulation between sonorant phonemes depends on (1) the phonetic context, (2) the phoneme durations, and (3) the spectral rate-of-change of the phonetic transitions (Lindblom, 1963; Moon and Lindblom, 1994). As was illustrated in Chapter 4, both vowel durations and the spectral rate-of-change generally increase with the linguistic prominence of the syllable in which a vowel occurs. However, a speaker's articulation effort can also change independently from the speaking rate, for example to adopt a fast, *hyper*-articulated style, or a slow, *hypo*-articulated style. The independence between speaking rate and articulation effort is captured in the contextual model of articulation, by considering the spectral rate-of-change as a separate parameter.

In this chapter, we investigate whether the contextual model of articulation can be integrated in a concatenative synthesis system. Specifically, we investigate whether the spectral rate-of-change of acoustic units can be modified, according to the prosodic structure of a target utterance, and we explore the potential of such modifications to improve the intelligibility and naturalness of concatenated speech.

In Section 5.2, a method is described to modify the spectral dynamics of acoustic units, using a line spectral frequency (LSF) representation. The target LSF dynamics are predicted based on the results of the statistical analysis in Chapter 4. Then, high-quality speech is generated corresponding to the modified LSF spectra, using a sinusoidal + all-pole parameterization of the acoustic units.

In Section 5.3, the proposed method is integrated in a Text-To-Speech (TTS) synthesis system. The naturalness of the proposed method is compared to standard concatenation techniques, using objective and subjective measures.

In a second experiment, we investigate the potential of the proposed method to improve the intelligibility of concatenated speech. Silverman *et al.* (1990) and others have reported that the intelligibility of synthetic speech decreases for longer sentences and paragraphs, likely because several acoustic-prosodic cues are missing compared to natural speech. In Section 5.4, we investigate whether the perception of syllabic stress in concatenative synthesis can be improved using the proposed method. The experiment is a first step at determining whether control of acoustic-prosodic effects, in addition to duration and pitch, can improve the intelligibility of concatenated speech.

5.2 Control of degree of articulation in concatenative speech synthesis

In this section, a method is proposed to integrate the contextual model of articulation in a concatenative synthesis framework. First, we describe a technique to modify the spectral shape of acoustic units, corresponding to a set of target spectral dynamics. The target dynamics are predicted based on the effects of prosodic factors analyzed in Chapter 4. Then, the parameterization of the speech signal is discussed, allowing spectral modification of acoustic units while maintaining the original speech quality.

5.2.1 Application of the contextual model of articulation

According to the contextual model of articulation, the degree of articulation of sonorant phonemes depends both on the phoneme durations and on the spectral rate-of-change, which reflects the speaker's articulation effort. This model can be integrated in a concatenative synthesis system, by predicting the spectral rate-of-change of speech segments in a target context, and modifying the spectral shape of the acoustic units accordingly. For a given acoustic unit and given phoneme durations, different phonetic targets are reached depending on the allowed spectral rate-of-change.

One way to control the spectral dynamics of acoustic units is to employ non-uniform time-warping. For example, if the spectral rate-of-change of a unit should decrease by 50%, this could be achieved by doubling the duration of phonetic transition regions inside the unit, and shortening stationary regions, such as at the center of a phoneme. However, this technique requires segmentation of acoustic units into stationary and transition regions, which makes the approach sensitive to alignment errors. Also, no vowel reduction or target undershoot could be achieved without removing speech in phonetic transition regions, which may lead to increased spectral mismatch.

Rather than using time-warping, we represent the acoustic units by a set of spectral trajectories related to the resonant frequencies of the vocal tract. The rate-of-change of these trajectories can be modified, for example using linear filtering. However, a filtering approach may not be optimal if the target rate-of-change, and hence the filter coefficients, varies continuously throughout an utterance, depending on the phonetic and prosodic context. Instead, the target trajectories can be computed as the most *likely* trajectories, given a set of static and dynamic constraints. This approach is explained in more detail below, for different spectral representations.

Formant trajectories

If a speech segment is represented by formant trajectories, where $f_{i,j}$ represents the jth formant at time points $i = 1..N$, then new trajectories $x_{i,j}$ can be defined that are close to $f_{i,j}$, and satisfy a set of dynamic constraints, denoted as $\Delta f_{i,j}$. The new trajectories are found by minimizing the following cost function with respect to $x_{i,j}$:

$$E = \sum_{i=1}^{N} (x_{i,j} - f_{i,j})^2 + D \sum_{i=1}^{N-1} (\Delta x_{i,j} - \Delta f_{i,j})^2 , \tag{5.1}$$

where D is a weighting factor determining the relative importance of the static versus the dynamic constraints. This cost function was first proposed by Plumpe *et al.* (1998), to smooth spectral trajectories using statistically predicted $\Delta f_{i,j}$ parameters, in an HMM framework. The trajectories $x_{i,j}$ can be shown to be the most likely trajectories satisfying the static and dynamics constraints, if $f_{i,j}$ and $\Delta f_{i,j}$ are normally distributed.

To modify the spectral rate-of-change, the dynamic constraints $\Delta f_{i,j}$ are specified as a scaled version of the local rate-of-change of the original formant trajectories:

$$\Delta f_{i,j} = k_i \Delta f_{i,j}^{orig}, \tag{5.2}$$

where k_i is a scaling function. If k_i is set to a constant k smaller than 1, Equation 5.2 expresses a uniform reduction in the spectral rate-of-change, corresponding to a constant reduction in the articulation effort.

In Figure 5.1, two stylized formant trajectories are shown along with their modified versions, when k is varied between 1, 0.6, and 0.4. Figure 5.1(a) represents a prototypical case of vowel reduction for one formant, where the difference between the trajectories at the midpoints corresponds to the amount of *target undershoot* (Lindblom, 1963). Figure 5.1(b) does not demonstrate target undershoot, but represents different degrees of co-articulation

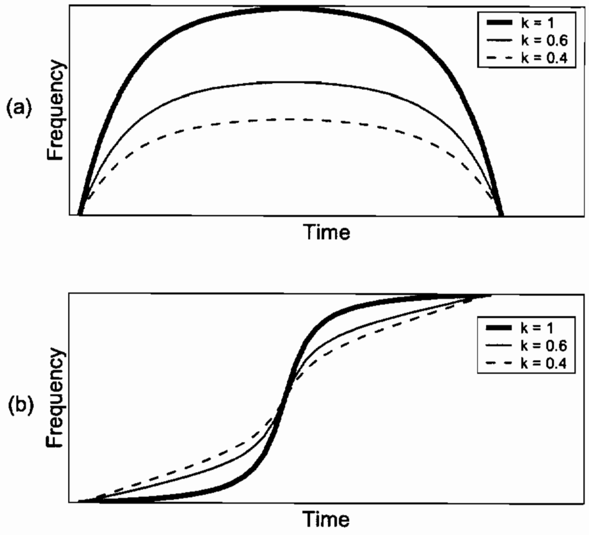

Figure 5.1: Stylized formant trajectories and reduced versions obtained using Equation 5.1.

between two sonorant phonemes for a single formant, resulting in less abrupt acoustic transitions.

For the stylized trajectories in Figure 5.1, $\Delta f_{i,j}^{orig}$ was set to $f_{i+1,j} - f_{i,j}$, and the weighting factor D was chosen such that the dynamic constraints $\Delta f_{i,j}$ were imposed most accurately. A value of $D = 40$ was selected, as the trajectory shapes changed very little when D was increased further. Specification of the parameters k_i, $\Delta f_{i,j}^{orig}$, and D in the context of an automated speech synthesis system is discussed below.

The endpoints of the trajectories in Figure 5.1 were "frozen", by setting $x_{i,j} = f_{i,j}$ for $i = 1, N$ in Equation 5.2. This is also motivated in the next section. Note that to create the reduced formant trajectories, no new vowel targets needed to be identified, nor was the vowel segmented into stationary and transition regions.

Line spectral frequency (LSF) trajectories

In concatenative synthesis, problems with formant detection prevent direct manipulation of the formant trajectories. Instead, the speech spectrum can be represented using line spectral frequencies (LSF), which are obtained automatically from LPC parameters. In

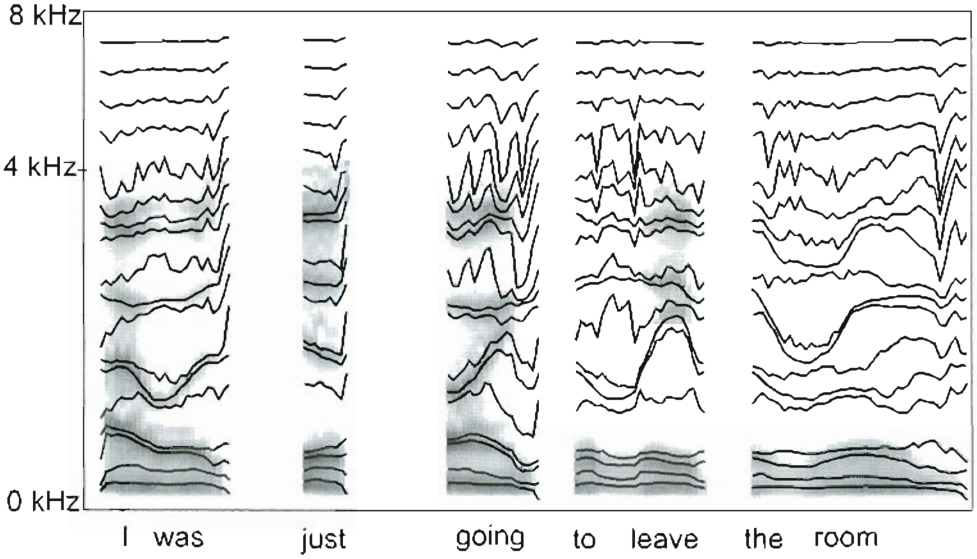

Figure 5.2: Line Spectral Frequency trajectories overlaid on pitch-synchronous, mel-warped, spectrogram.

this work, the LSF's are computed from a sinusoidal + all-pole representation of speech, as is discussed in Section 5.2.3.

In Figure 5.2, LSF trajectories are overlaid on a pitch-synchronous, mel-warped, spectrogram. Because the distance between two closely placed LSF's corresponds to the bandwidth of an underlying LPC pole, the trajectories of close LSF pairs frequently coincide with formant trajectories, as can be verified in Figure 5.2. Hence, the LSF's are a suitable representation to enable modification of the formant dynamics. Another advantage of the LSF representation is that modified LSF parameters can be converted back to LPC filters that are stable as long as the ordering of the LSF's is preserved (Soong and Juang, 1984).

Equation 5.1 can be applied to control the spectral rate-of-change of the LSF trajectories, similar to the trajectories in Figure 5.1. The modified LSF trajectories can then be converted back to speech using the sinusoidal + all-pole model discussed in Section 5.2.3. However, application of Equation 5.1 to each LSF trajectory independently does not preserve the distance between closely placed LSF's, resulting in formant widening or in the insertion of spurious high-energy peaks in the speech spectrum. Both artifacts degrade the perceptual quality of the resulting speech signal.

The distance between LSF pairs can be controlled by extending Equation 5.1 with a third term:

$$
\begin{aligned}
E &= \sum_{i=1}^{N}\sum_{j=1}^{M}(x_{i,j} - f_{i,j})^2 \\
&\quad + \sum_{i=1}^{N-1}\sum_{j=1}^{M} D_1(i,j)(\Delta^i x_{i,j} - \Delta^i f_{i,j})^2 \\
&\quad + \sum_{i=1}^{N}\sum_{j=1}^{M-1} D_2(i,j)(\Delta^j x_{i,j} - \Delta^j f_{i,j})^2. \qquad (5.3)
\end{aligned}
$$

where

$f_{i,j}$	initial LSF value at time i in trajectory j
$\Delta^i f_{i,j}$	desired time-derivative at i, j
$\Delta^j f_{i,j}$	desired LSF distance, normally set to $f_{i,j+1} - f_{i,j}$
$x_{i,j}$	new LSF value at time i in trajectory j
$\Delta^i x_{i,j}$	time-derivative of $x_{i,j}$
$\Delta^j x_{i,j}$	LSF distance between $x_{i,j+1}$ and $x_{i,j}$
N	number of time points in each LSF trajectory
M	number of LSF trajectories, corresponding to the LPC model order.

If $\Delta^i x_{i,j}$ and $\Delta^j x_{i,j}$ are linear expressions of $x_{i,j}$, Equation 5.3 is easily minimized, by setting $dE/dx_{i,j} = 0$, and solving the resulting set of linear equations, using linear regression.

The weights $D_2(i,j)$ are specified so that the distance between closely placed LSF's is maintained, in order to preserve the bandwidths of the underlying formants. This can be achieved by setting $D_2(i,j)$ to a value inversely proportional to $|\Delta^j f_{i,j}|^2$. Similarly, the weights $D_1(i,j)$ are made proportional to $\Delta^i f_{i,j}$, so that the target spectral dynamics are imposed more accurately for rapidly changing LSF's, which often coincide with formant trajectories (see Figure 5.2).

Therefore, the following formula was used for $D_1(i,j)$ and $D_2(i,j)$:

$$D_p(i,j) = a_p + b_p\,|\Delta^p f_{i,j}|^{c_p}, \qquad p = 1, 2 \qquad (5.4)$$

where $c_1 = 1$ and $c_2 = -2$. The system parameters a_p and b_p were determined by generating a set of test sentences using different values for a_p and b_p, and retaining the values that produced the best perceptual results, through informal listening tests. In future work, the parameters should be further optimized through tests with multiple listeners, or using a suitable objective quality measure.

In the present work, the contextual model of articulation is applied only in sonorant regions of speech. While the articulatory movements in non-sonorant speech may be governed by similar dynamics as in sonorant speech, the resonant frequencies of the vocal tract are usually not observable in the spectrum of non-sonorants, or must be separated from aperiodic signal components. Therefore, the trajectories $f_{i,j}$ are "frozen" at the sonorant boundaries, by replacing the relevant $x_{i,j}$ by $f_{i,j}$ in Equation 5.3. An alternative boundary condition is to set $D_1(i,j) = 0$ at the sonorant boundaries, allowing trajectory endpoints to change more freely. However, this approach may result in spectral discontinuities between sonorant and non-sonorant segments, and ignores the fact that the (hidden) articulatory dynamics in non-sonorant speech may still affect the trajectories in sonorant speech.

The cost function in Equation 5.3 was first introduced in Chapter 3 in the context of *unit fusion*. There, $\Delta^i f_{i,j}$ was defined to reduce spectral mismatch at concatenation points, by utilizing the spectral dynamics of external "fusion" units. Here, the objective is to improve the acoustic-prosodic characteristics of concatenated units, by specifying spectral dynamics $\Delta^i f_{i,j}$ that reduce mismatch between the degree of articulation of acoustic units in the database and in the target prosodic context. This is discussed in the next section.

5.2.2 Effects of prosodic factors on $\Delta^i f_{i,j}$

In Chapter 4, a numerical model was developed that describes the spectral rate-of-change *RoC* of a phonetic transition as the product of five independent factors:

$$RoC = K_p \; \alpha(S) \; \beta(A) \; \gamma(P) \; \delta(Y). \tag{5.5}$$

K_p is determined by the transition identity. The functions α to δ map the discrete level of four prosodic factors, i.e., stress (S), accent (A), word position (P), and speaking style (Y), to a numerical value. Hence, the functions α to δ modify the spectral rate-of-change

Table 5.1: Numerical values for prosodic factors, when $\alpha(S{=}0)$, $\beta(A{=}0)$, $\gamma(P{=}\,\textit{final})$, $\delta(Y{=}\,\textit{clear})$, and $\epsilon(W{=}\,\textit{func})$ are normalized to 1.

	Onset	Diphthong	Coda
$\alpha(S{=}1)$	1.18	1.01	0.97
$\beta(A{=}1)$	1.11	1.05	1.07
$\gamma(P{=}\,\textit{medial})$	1.07	1.05	1.19
$\delta(Y{=}\,\textit{fast})$	0.98	1.13	1.10
$\delta(Y{=}\,\textit{relaxed})$	0.70	1.06	0.92
$\epsilon(W{=}\,\textit{cont})$	1.10	1.10	1.10

RoC of a phonetic transition, depending on the prosodic context. The numerical values were estimated for liquid-vowel, diphthong, and vowel-liquid transitions, based on an experimental database, and are summarized in Table 5.1. Since the liquid-vowel transitions occured in syllable onsets and the vowel-liquid transitions occured in syllable coda's, the three transition types are referred to as *onset*, *diphthong*, and *coda* transitions.

The model in Equation 5.5 can be exploited to control the degree of articulation of concatenated speech, by modifying the rate-of-change of acoustic units depending on the target prosodic context. The target spectral dynamics are defined as a scaled version of the original spectral dynamics of an acoustic unit,

$$\Delta^i f_{i,j} = k_i \Delta^i f_{i,j}^{orig}, \tag{5.6}$$

where k_i is a scaling function corresponding to the ratio of the target prosodic factors of an acoustic unit and the original prosodic factors,

$$k_i = \frac{\alpha(S^t)\ \beta(A^t)\ \gamma(P^t)\ \delta(Y^t)\ \epsilon(W^t)}{\alpha(S^o)\ \beta(A^o)\ \gamma(P^o)\ \delta(Y^o)\ \epsilon(W^o)}. \tag{5.7}$$

This expression for k_i was found by equating $\Delta^i f_{i,j}$ and $\Delta^i f_{i,j}^{orig}$ to *RoC* in Equation 5.5, and substituting K_p. The superscripts t and o refer to the levels of the target and the original prosodic factors, respectively; ϵ is a prosodic factor related to word class, as is motivated below.

The parameter k_i can be interpreted as a gain factor, increasing the spectral rate-of-change of an acoustic unit when the target utterance requires a higher degree of articulation than the original utterance, or, conversely, decreasing the spectral rate-of-change of a unit when the target utterance corresponds to a less prominent context than the original utterance. For example, if an acoustic unit occurs at the onset of a stressed syllable in the databse, but is needed in an unstressed syllable during synthesis, the spectral rate-of-change is scaled by $k_i = 1/1.18 = .85$ (see Table 5.1), reducing the rate-of-change by 15%.

Unfortunately, the numerical data available in Table 5.1 are sparse compared to the information needed to compute k_i for synthesis of unrestricted text. The prosodic factors studied in Chapter 4 do not include all the factors known to have an effect on vowel quality, such as the syntactic category of a word (van Bergem, 1993), or word length (Moon and Lindblom, 1994). Furthermore, only two or three levels were studied per prosodic factor, ignoring effects of, for example, secondary syllabic stress, different pitch accents, and intermediate phrase boundaries. Therefore, the prediction of k_i described in this section is only a first approach to integrate the results from Chapter 4 in a concatenative synthesis system. A more robust way to predict k_i would be to use statistical methods to deconfound factors in a corpus of fluent speech, as proposed by van Santen (1992) in the context of duration modeling.

In Chapter 4, significant differences were found between the prosodic factors for onset, diphthong, and coda transitions, indicating that the scaling function k_i is not constant throughout each syllable. Therefore, we use Equation 5.7 to define k_i at the start and end point of each vowel, taking the numerical values α to δ from the *onset* column in Table 5.1 at the start of a vowel, and from the *coda* column at the end of a vowel. Then, k_i is linearly interpolated for intermediate time-points i in each vowel. In liquids and glides, k_i depends on the syllabic structure. In syllable codas, k_i is set to the endpoint value of the preceding vowel. In syllable onsets, the endpoints of k_i are set to the value of the neighboring sonorant, or to 1 at the transition with non-sonorants. Interpolation is used for the intermediate liquid/glide time-points.

The linear interpolation scheme assumes that k_i changes gradually over the course of a syllable. The validity of this assumption could be tested by measuring the prosodic effects at different points in a syllable, while issues of time-normalization and measurement accuracy for low RoC values would need to be resolved. One set of such measurements is provided by the diphthong data in Table 5.1. The numerical values are comparable to those for coda transitions, which might be expected as the diphthong transitions occur often close to the syllable coda. However, the analysis in Chapter 4 showed that the effects of the prosodic factors for diphthongs were estimated less reliably, possibly because the effects of the prosodic factors on the diphthong transitions were dominated by other articulatory constraints. Therefore, we decided to use only the onset and coda values in Table 5.1 to compute k_i.

In addition to the prosodic factors studied in Chapter 4, a factor based on word class, ϵ, was included in Equation 5.7. This factor was normalized to 1 in function words, and was set to 1.1 otherwise, so that the spectral rate-of-change (i.e., articulation effort) is relatively lower in function words than in other word classes (van Bergem, 1993).

In Equation 5.6, $\Delta^i f_{i,j}^{orig}$ represents a smooth time-derivative of $f_{i,j}$. Usually, the synthesis time-points i of a target utterance are mapped to analysis time-points $w(i)$ in the acoustic units, to perform time- and pitch-scale modifications. To avoid effects of $w(i)$ on $\Delta^i f_{i,j}^{orig}$, the time-derivative is based on the LSF representation of the acoustic units prior to time-scale modification. In the present work, the derivative was computed using a 6-point FIR filter, corresponding to a low-pass smoothing operation convolved by the first order difference. The LSF derivative was multiplied by the ratio of the original and the target fundamental frequency to account for pitch-scale modifications.

5.2.3 Spectral modification

The LSF trajectories $x_{i,j}$, obtained by minimizing Equation 5.3, must be transformed back to a speech signal with a perceptual quality comparable to that of unmodified acoustic units. A technique to achieve this was described in Chapter 2, and is briefly reviewed here. Figure 5.3 summarizes the analysis/synthesis process.

During the analysis stage, each pitch-synchronous two-period speech frame $s_i[n]$ is represented by a harmonic estimate,

$$\hat{s}_i[n] = \sum_{k=-L}^{L} a_{i,k} \exp(jk\omega_0 n), \tag{5.8}$$

$$\begin{tikzcd}
\text{ANALYSIS} & & \text{SYNTHESIS} & \\
s_i[n] \arrow[d] & & s'_i[n] & \text{time-domain signal} \\
a_{i,k} \arrow[r] \arrow[d] & r_{i,k} \arrow[r, "\text{warp(k)}"] & a'_{i,k} \arrow[u] & \text{sinusoidal model} \\
S(\omega) \arrow[ur] \arrow[d] & & S'(\omega) \arrow[u] & \text{all-pole model} \\
f_{i,j} \arrow[rr, "\text{Equation 3}"] & & x_{i,j} \arrow[u] & \text{LSF coefficients}
\end{tikzcd}$$

Figure 5.3: Analysis/Synthesis based on sinusoidal + all-pole model of speech.

where ω_0 is the fundamental frequency, L represents the number of harmonics, and $a_{i,k} = B_{i,k} \exp(j\phi_{i,k})$ are the complex sinusoidal amplitudes. The sinusoidal parameters are found by solving a complex regression, which minimizes the squared error between $s_i[n]$ and $\hat{s}_i[n]$ (Stylianou, 2001). From $a_{i,k}$, speech can be reconstructed that is practically indistinguishable from $s_i[n]$, by evaluating Equation 5.8 pitch-synchronously and applying overlap-add (McAulay and Quatieri, 1986; Stylianou, 2001).

After sinusoidal analysis, an all-pole model is fit to the sinusoidal parameters. First, the power spectrum $|a_{i,k}|^2$ is mel-warped to improve the resolution of the all-pole model towards lower frequencies, corresponding to the frequency resolution of human hearing. To reduce bias of the all-pole spectrum towards the harmonic frequencies, the power spectrum is upsampled, using cubic interpolation in the log domain (Hermansky *et al.*, 1984). Then, correlation coefficients are obtained by an Inverse Discrete Fourier Transform (IDFT) of the power spectrum. Application of the Levinson-Durbin algorithm to the correlation coefficients yields the all-pole (LPC) parameters. In the present work, an all-pole model of order $M = 18$ was chosen for 16 kHz speech. The LPC parameters define a smooth all-pole spectrum $S_i(\omega) = \sigma^2/A(e^{-j\omega})$, which is converted to LSF's $f_{i,j}$ by finding the roots of the polynomials $P(z) = A(z) + z^{-(M+1)}A(z^{-1})$ and $Q(z) = A(z) - z^{-(M+1)}A(z^{-1})$ (Itakura, 1975).

In the synthesis stage, waveforms $s'_i[n]$ must be determined corresponding to $x_{i,j}$. As illustrated on the right hand side of Figure 5.3, the LSF values $x_{i,j}$ corresponding to a synthesis frame are first converted to LPC coefficients, defining a smooth all-pole spectrum $S'_i(\omega)$. The sinusoidal parameters $a'_{i,k}$ can then be obtained by evaluating $S'(\omega)$ at the harmonics of the target fundamental frequency ω'_0, i.e., $a'_{i,k} = S'_i(k\omega'_0)$. However, the all-pole envelope does not model some perceptually important characteristics of speech, for example by imposing a minimum-phase spectrum (Quatieri and McAulay, 1987). Choosing $a'_{i,k} = S'_i(k\omega'_0)$ results in speech with a quality comparable to that of impulse-excited LPC.

A new method to predict $a'_{i,k}$ was proposed in Chapter 2 (also Wouters and Macon, 2001a). The method is based on computing a spectral residual $r_{i,k} = a_{i,k}/S(k\omega_0)$ during speech analysis, and multiplying the target all-pole spectrum $S'(\omega)$ by a frequency-warped version of $r_{i,k}$. Resampling of the resulting spectrum produces the target sinusoidal parameters $a'_{i,k}$. The frequency-warping is a monotonic, piece-wise linear function, which maps dominant poles of $S(\omega)$ to dominant poles of $S'(\omega)$. Hence, residual characteristics $r_{i,k}$ around the spectral peaks or in the valleys of $S(\omega)$ are transferred to equivalent regions in $S'(\omega)$. As a result, perceptually important characteristics of $a_{i,k}$ are preserved in $a'_{i,k}$, while large residual values in the valleys of $S(\omega)$ do not degrade the shape of the target formant peaks, as occurs in modifications based on residual-excited LPC (RELP).

From $a'_{i,k}$, $s'_i[n]$ is computed by applying Equation 5.8. A gain factor is applied to $s'_i[n]$ to ensure that the RMS energy of the analysis frame $s_i[n]$ is preserved. Alternatively, the log RMS energy contour of the acoustic units can be controlled similar to an LSF trajectory, in order to reduce the intensity of linguistically less prominent vowels, and to reduce energy mismatches at concatenation points.

5.3 Text-To-Speech experiment

5.3.1 Material

To evaluate the method in Section 5.2, eighty English sentences were generated, using three different synthesis methods. The acoustic units were selected from a diphone table, and had been recorded in the context of small word groups or nonsense words. The speaker was the same person that recorded the database in Chapter 4. Hence, control of spectral dynamics based on the data in Table 5.1 should lead to a more accurate realization of this speaker's degree of articulation in fluent speech.

The sentences were synthesized using phoneme durations and fundamental frequency contours derived from natural sentences. Twenty sentences corresponded to short phrases recorded by the database speaker. Sixty other sentences were recorded by a different speaker, using a more lively speaking style. The "lively" sentences included several long words and function words, which introduce large variations in articulation effort in natural speech, and were expected to highlight quality improvements when controlling the degree of articulation in concatenated speech.

The sentences were synthesized using three different methods. These methods differed in their approach towards resolving spectral mismatch, either between two concatenated units (i.e., "concatenation errors"), or between the prosodic context of a given unit and the target utterance (i.e., "prosodic target errors"). The spectral structure of the acoustic units was modified in two of the three synthesis methods. Then, the units were combined using pitch-synchronous overlap-add to generate the final utterance.

The first synthesis method was a time-domain method. Time- and pitch-scale modifications were achieved using TD-PSOLA (Moulines and Charpentier, 1990), and units were concatenated pitch-synchronously. No spectral modifications were performed. This method was denoted as "TD".

In the second method, spectral modification was enabled at the concatenation points. Concatenation mismatch between the LSF parameters of two units was spread linearly over an interpolation region, as described by Dutoit and Leich (1994). The interpolation region was limited to the phoneme in which a concatenation occured, and it included at most 100 ms on each side of the concatenation point. This method was denoted as "LIN".

In the third method, the effects of prosodic factors were taken into account by specifying a scaling function k_i, as described in Section 5.2.2, and modifying the spectral shape of the acoustic units accordingly. Since the acoustic units consisted of diphones, spoken in a prosodically neutral context, their spectral transitions were assumed to be maximally

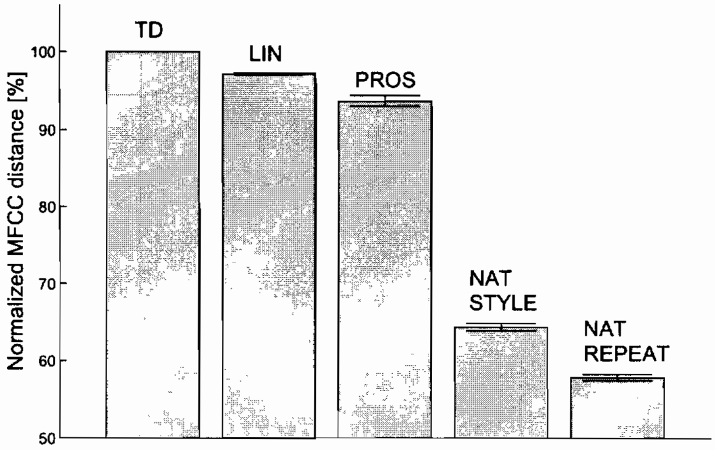

Figure 5.4: Normalized MFCC distance between copy-prosody sentences and natural examples, for time-domain concatenation method (TD), linear interpolation of LSF's (LIN), control of prosodic effects (PROS).

articulated, i.e., S^o=1, A^o=1, P^o=*medial*, Y^o=*clear*, W^o=*func* in Equation 5.7. The target prosodic factors were predicted automatically by the TTS engine, except for the speaking style, Y^t, which was set to *relaxed* for fast sentences (see below), and to *clear* otherwise. The third synthesis method was denoted as "PROS".

In the three synthesis methods, the log RMS energy of the synthesis frames was treated similar to the LSF trajectories. Hence, in LIN, intensity mismatch between concatenated units was reduced using linear interpolation. In PROS, the time-dynamics of the energy contour were controlled to reduce acoustic-prosodic mismatches.

5.3.2 Results

The performance of the three synthesis methods was compared objectively and subjectively. As an objective measure, Mel-frequency cepstral (MFCC) distances were computed between the synthesized utterances and their natural examples, for the twenty sentences recorded by the database speaker. MFCC distances have been shown to correlate moderately well with perceptual distance measurements (e.g., $\rho = .66$ in Wouters and Macon

(1998), $\rho = .67$ in Chen and Campbell (1999)), although not in the case of spectral discontinuities, as reported by (Klabbers and Veldhuis, 1998). MFCC features are also commonly used in automatic speech recognition and to guide unit selection in speech synthesis (Hunt and Black, 1996).

The MFCC feature vectors were computed pitch-synchronously in sonorant phonemes, based on the sinusoidal + all-pole representation described in Section 5.2.3. The cepstral means were subtracted per sentence, to reduce the effect of differences in the recording conditions for the acoustic units and the example sentences. For each synthesis method m and sentence s, the objective distance $D(m, s)$ was defined as the root-mean-square Euclidean distance between the MFCC feature vectors of the synthesized utterance and of the natural example. $D(m, s)$ was divided by $D(TD, s)$, i.e., the distance obtained for the time-domain synthesis method, to normalize for sentence effects. The normalized distances were then averaged per synthesis method m.

The results are shown in Figure 5.4. The three synthesis methods progressively decrease the distance between the concatenated and the natural utterances. These distances can be compared to the *intra-speaker variability*, or the average spectral distance between phonemically identical units recorded by the database speaker. The intra-speaker variability was computed using the phonetic transition regions described in Chapter 4. There, each sentence containing a particular phonetic transition was repeated nine times, i.e., three times for each of three speaking styles. The average distance between repetitions of the phonetic transition across speaking styles was 63% of the average distance $\overline{D(TD, s)}$. Within the same speaking style, the average distance between repetitions was 56% of $\overline{D(TD, s)}$.

Figure 5.4 suggests that the proposed method has an effect equal in size to changes in speaking style in natural speech. However, the distance between concatenated speech and natural speech remains significantly larger than the intra-speaker variability. This can be attributed partly to the fact that the acoustic units in this experiment consisted of diphones, which are typically fully articulated and may have a different voice quality than is used in natural speech. The results show that the proposed method should be further improved to control the degree of articulation of diphone units, and that other acoustic-prosodic characteristics, such as aspects of voice quality, should also be modified.

Table 5.2: Illustration of complementary version design when three pairs of synthesis methods, denoted as A, B, and C, need to be compared. The columns correspond to sentences and the rows correspond to listeners.

	S1	S2	S3	
L1	A	B	C	...
L2	B	C	A	
L3	C	A	B	
	⋮			

While approaches based on selecting units from a large corpus of fluent speech may yield utterances closer to natural speech, we believe that the proposed method can further improve the naturalness of such utterances, by controlling the degree of articulation when prosodic target mismatch occurs.

The three synthesis methods were also evaluated perceptually using the Comparative Mean Opinion Score (CMOS) test (ITU P.800, 1996). Fifteen subjects listened to pairs of synthesized utterances, and indicated their preference on a 7-point scale, ranging from "A is Much Worse than B" (-3) to "A is Much Better than B" (+3). All listeners were fluent speakers of American English. They listened over headphones in a sound-insulated booth, and could play the utterances several times before selecting a score.

For each sentence, three pairs of synthesis methods needed to be evaluated, i.e., TD-LIN, LIN-PROS, and TD-PROS. A complementary version design was adopted, to limit the duration of the test and to avoid training effects (van Santen, 1993). According to this design, subsequent listeners hear different synthesis pairs for each sentence, but each pair is evaluated equally often per sentence across listeners, and each listener hears each pair equally often across sentences. An example is shown in Table 5.2, for the case where three pairs of synthesis methods need to be evaluated. The total number of sentences and listeners is also a multiple of 3. The ordering of the sentences and the synthesis methods within a pair is randomized per listener.

Thirty sentences were evaluated: nine "fast" sentences, nine "neutral" sentences, and twelve "lively" sentences. Additionally, one sentence for each group was presented as an

Table 5.3: Average perceptual scores of Comparative Mean Opinion Score (CMOS) test.

	LIN	PROS
TD	.50 ± .08	.96 ± .09
LIN		.43 ± .09

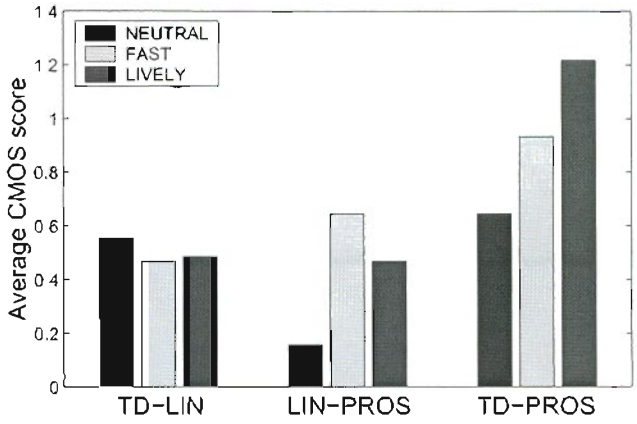

Figure 5.5: Average CMOS scores for time-domain concatenation method (TD), linear interpolation method (LIN), and proposed method (PROS), for "neutral", "fast", and "lively" sentences.

example to the listeners at the beginning of each test session.

The sentences were generated as follows. Of the twenty sentences recorded by the database speaker, the phoneme durations were shortened by 20% for ten sentences, to generate "fast" speech. The remaining ten sentences were labeled as "neutral" speech. From the sixty "lively" sentences, thirteen (i.e., twelve test sentences plus one example) were selected by finding the sentences for which the difference between the LIN and the PROS version was largest, using an objective distance measure as well as informal subjective comparisons. This preselection step was motivated by the fact that the proposed method does not always significantly alter the concatenated units, for example when there are no large formant movements or prosodic mismatches.

The thirty sentences were evaluated using the complementary version design described earlier, such that each listener evaluated one utterance pair per sentence, resulting in a

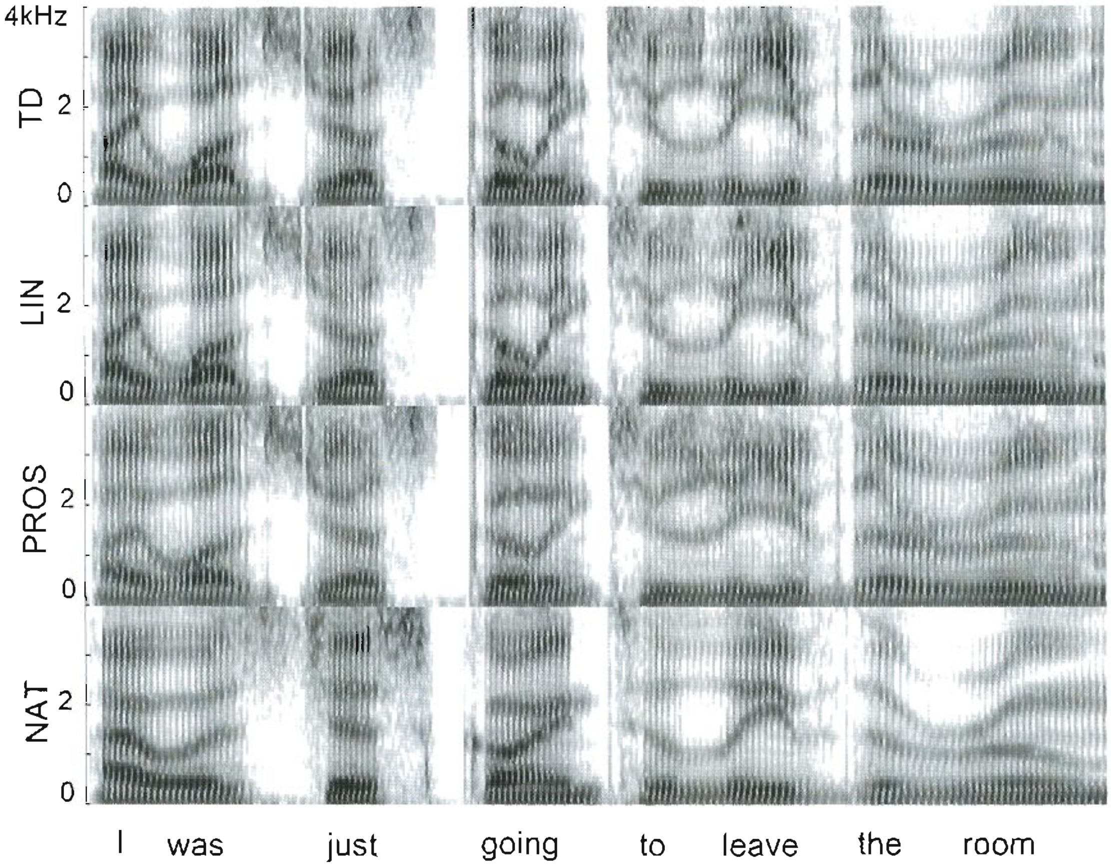

Figure 5.6: Spectrograms corresponding to time-domain concatenation method (TD), linear interpolation of LSF's (LIN), control of prosodic effects (PROS).

total of thirty judgements. The average perceptual scores are shown in Table 5.3, together with their 5% confidence intervals. The results show that the linear interpolation method (LIN) improved the perceptual quality compared to the time-domain method (TD), while the proposed method (PROS) performed better than either TD or LIN. Analysis of the perceptual results per sentence group showed that the improvement of PROS compared to LIN was larger for the "lively" and "fast" sentences than for the "neutral" sentences. This is illustrated in Figure 5.5.

In Figure 5.6, spectrograms corresponding to one of the test sentences are shown for the three synthesis methods. Comparison between TD and LIN shows that linear smoothing can be effective when the specified interpolation region is appropriate, such as in the words "the" and "room". In the PROS method, the dynamic constraints have reduced the spectral rate-of-change in the words "I was" and "leave", while smooth trajectories are obtained for the short acoustic units in "going".

The utterances discussed in this section are available at `http://cslu.cse.ogi.edu/tts/demos/jasa01`.

5.4 Stress perception experiment

Although current synthesizers attain high intelligibility scores for isolated words, long sentences or paragraphs are often harder to understand, or impose a higher cognitive load than natural speech (Silverman and Morgan, 1990; van Santen, 1997; Delogu *et al.*, 1998). In this section, an experiment is described to evaluate the potential of the proposed method to improve the perception of syllabic stress in concatenated speech. The experiment is a first step at determining whether control of acoustic-prosodic effects in addition to time and pitch scale modifications can increase the intelligibility of concatenated speech.

In English, vowel quality is but one of the acoustic correlates of syllabic stress, next to pitch, duration, intensity, and spectral balance (e.g., Sluijter and van Heuven, 1996; Waibel, 1986; Lindblom, 1963). The present experiment tests the effect of vowel quality on stress perception, for given values of pitch, phoneme duration, intensity, and spectral balance. If the experiment shows that stress perception indeed depends on vowel quality, then the proposed method can be used to improve the perception of stress in concatenative speech, by controlling the spectral shape of acoustic units in addition to other variables.

5.4.1 Material

The test material consisted of thirty sentences, synthesized in eight different ways. The sentences were of the form "Please produce W for him", where W is a two-syllable word that can carry lexical stress on either syllable in English. For example, the word "digest" carries lexical stress on the first syllable σ_1 when it is a noun, and on the second syllable σ_2 when it is a verb. The two-syllable words were taken from the study in Chapter 4. The objective of the present experiment was to spectrally reduce σ_1 or σ_2 using the proposed method, and to study whether stress perception shifted away from the reduced syllable. Sluijter and van Heuven (1997) used a similar approach to evaluate the effect of spectral balance on stress perception, for the nonsense word "nana".

Table 5.4: Utterances generated for each two-syllable word W. σ_1 and σ_2 refer to the first and second syllable of W, respectively. The rows indicate whether the pitch contour is normalized in W, whether the phoneme durations correspond to a stressed-unstressed pattern ($U1$) or unstressed-stressed pattern ($U8$), and whether σ_1 or σ_2 is spectrally reduced. The bottom row contains the results of the stress perception experiment, reported as the frequency $p(\sigma_1)$ with which the first syllable is perceived as stressed.

	U1	U2	U3	U4	U5	U6	U7	U8
norm F0		x	x	x	x	x	x	
σ_1 short					x	x	x	x
σ_2 short	x	x	x	x				
σ_1 reduced				x		x	x	x
σ_2 reduced	x	x	x		x			
$p(\sigma_1)$ [%]	65	51	47	39	32	27	29	19

Eight utterances were generated per two-syllable word, as summarized in Table 5.4. The utterances were based on two natural recordings, which are denoted as $U1$ and $U8$ in Table 5.4. The speaker assigned syllabic stress to σ_1 in $U1$, and to σ_2 in $U8$. The speaker placed a nuclear pitch accent on the word "again", which followed the carrier phrase. This was done to avoid large pitch movements in σ_1 or σ_2, which would dominate stress perception and limit the effects of further stress manipulations. The accented word "again" was not included in the utterances used for the experiment.

A straight line was fit to the original pitch contour of $U1$, to specify a target $F0$ without local pitch movements in W. A new utterance, denoted as $U2$, was generated with this target $F0$, by applying PSOLA modifications (Moulines and Charpentier, 1990) to $U1$. Similarly, a pitch-normalized utterance, denoted as $U7$, was generated from $U8$.

A concatenated utterance, denoted as U', was generated by replacing the unstressed syllable σ_2 in $U1$ by the stressed version of σ_2 from $U8$. Hence, U' contained a sequence of two naturally stressed syllables, as can occur during unit selection when a desired unstressed unit is not found in the unit database.

U' was subjected to modifications of duration, pitch, and spectral shape, as follows. In $U3$ and $U4$, the phoneme durations of σ_2 in U' were changed to the unstressed (i.e., short) durations of σ_2 in $U1$. In $U5$ and $U6$, the phoneme durations of σ_1 in U' were changed to the unstressed durations of σ_1 in $U8$. The pitch contours were normalized as for $U2$ or $U7$. Then, σ_2 was spectrally reduced in $U3$ and $U5$, using the proposed method. Similarly, σ_1 was reduced in $U4$ and $U6$. The scaling values $\alpha(S{=}1)$ in Table 5.1 were raised by 10% to increase the amount of vowel reduction. The parameter $D_1(i,j)$ in Equation 5.3 was also increased, so that the target spectral dynamics were imposed more accurately.

In the natural utterances $U1$ and $U2$, spectral reduction of σ_2 is the result of natural articulation processes. Similarly, σ_1 is naturally reduced in $U7$ and $U8$.

5.4.2 Results

The utterances were presented to sixteen listeners, who were native speakers of American English, without known hearing problems. Each listener heard four utterances per word W, or 120 utterances, using the complementary version design described in Section 5.3. The utterances were presented over headphones in a sound-insulated booth. The participants could listen several times to each utterance, by manipulating a graphical interface. Then, they were asked to classify W, by selecting one of two displayed words, corresponding to alternative stress patterns of W, e.g. "digest" and "digest". After making a selection, listeners proceeded to the next utterance.

The resulting classification scores, averaged over the sixteen listeners, are shown in the final column of Table 5.4. The scores correspond to the frequency $p(\sigma_1)$ with which the first syllable was perceived as stressed for a particular utterance type. In Figure 5.7, the classification scores are presented graphically, along with the 5% confidence intervals. For the natural utterances U1 and U8, respectively the first or second syllable was intended to be stressed. The classification scores of 65.4% and 19.2%, indicate that the intended stress pattern was not always perceived correctly, which may be attributed to the absence of a strong pitch accent and the lack of semantic information. Although no strong pitch

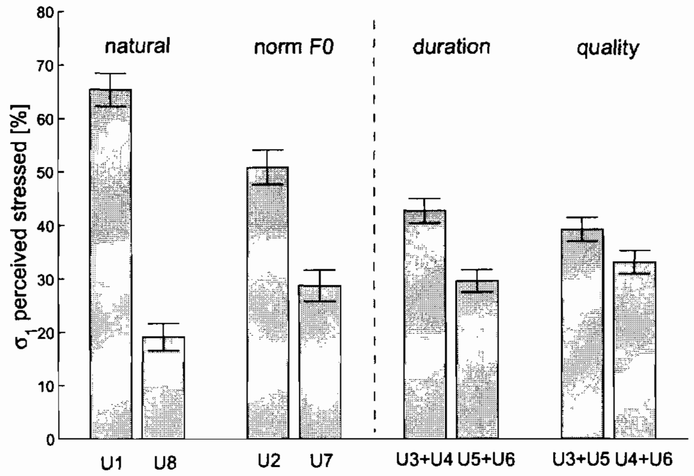

Figure 5.7: Classification of first syllable as stressed for (a) natural utterances $U1$ and $U8$, (b) natural utterances with normalized pitch $U2$ and $U7$, (c) concatenated utterances grouped per durational stress pattern, and (d) concatenated utterances grouped per spectrally reduced syllable.

accents occured in W, the classification accuracy further decreased when pitch movements were normalized in $U2$ and $U7$.

For the utterances based on U', modifying the phoneme durations introduced an average difference of 13% in $p(\sigma_1)$ (p<.0001, double-sided t-test), while modifying the vowel quality introduced an average difference of 6% in $p(\sigma_1)$ (p=.05, double-sided t-test). The effects of duration and vowel quality modifications combined, so that when the second syllable is short *and* spectrally reduced, $p(\sigma_1)$ = 46.7% approaches the result for $U2$. When the first syllable is short *and* spectrally reduced, $p(\sigma_1)$ = 27.5% is slightly lower than the result for $U7$. Hence, the proposed method changes the degree of articulation of the concatenated units, so that the intended stress pattern is perceived more accurately.

The utterances discussed in this section are available at
`http://cslu.cse.ogi.edu/tts/demos/jasa01`.

5.5 Unit Fusion and Control of Degree of Articulation

In Chapter 3, unit fusion was introduced as a technique to reduce concatenation mismatch between acoustic units. The technique represents a "template-based" approach to determine the target spectral dynamics at concatenation points. On the other hand, in the present chapter we explored whether the degree of articulation of acoustic units can be modified based on the contextual model of articulation. The spectral dynamics of the original acoustic units were modified using a scaling function k_i, in order to model differences between the original and target prosodic context. This method was referred to as the "prosodic control" method.

In addition to reducing concatenation mismatch, application of unit fusion can also change the degree of articulation of acoustic units. This occurs because the dynamics of a fusion unit are applied in a region around a concatenation point (see Figure 3.1). If a fusion unit is selected that is appropriate for the target prosodic context, the target spectral dynamics specified using unit fusion may be similar to those specified using the prosodic control method. However, unit fusion by itself may not always control the degree of articulation successfully, for example when the spectral dynamics should change outside a concatenation region, or when the selected fusion unit does not represent the target prosodic context.

In the TTS experiment in Section 5.3, three synthesis methods were evaluated, i.e., time-domain concatenation (TD), linear interpolation (LIN), and the prosodic control method (PROS). These synthesis methods can be compared with unit fusion (FUS), and with unit fusion followed by prosodic control (FUS+PROS). In the last method, the spectral dynamics of the fusion units are first combined with the dynamics of the concatenation units using:

$$\Delta^i f_{i,j}^{orig} = \alpha_i \Delta^i f_{v(i),j}^{fusion} + (1 - \alpha_i) \Delta^i f_{w(i),j}^{concat}, \tag{5.9}$$

where $v(i)$ and $w(i)$ are warping functions that map synthesis time-points in the target utterance to time-points in the fusion and concatenation units, respectively (see Equation 3.2). The parameters $\Delta^i f_{i,j}^{orig}$ then form the input to Equation 5.6.

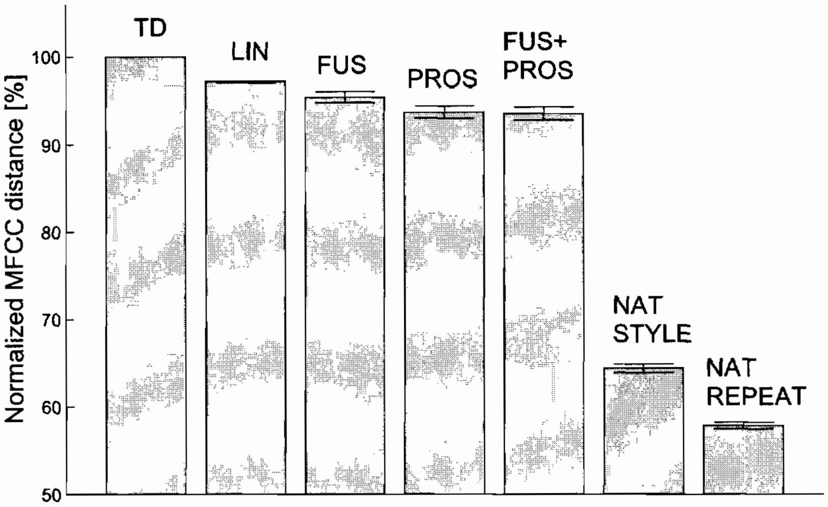

Figure 5.8: Relative objective distance between synthesized utterances and natural utterances, for different synthesis methods: time-domain (TD), linear interpolation (LIN), unit fusion (FUS), prosodic control (PROS), unit fusion followed by prosodic control (FUS+PROS).

The methods TD, LIN, FUS, PROS, and FUS+PROS were compared using the sentence material and the objective quality measure discussed in Section 5.3. The results are shown in Figure 5.8. The unit fusion technique produces utterances that are closer to natural speech than utterances produced by the linear interpolation method. However, combination of unit fusion and the prosodic control method does not improve the result compared to the prosodic control method by itself. This may be explained by the fact that the spectral dynamics predicted in the prosodic control method already impose consistent trajectories at concatenation points, which are not further improved using unit fusion.

Informal listening tests confirmed that utterances produced by the unit fusion method sounded more natural than those produced by the linear interpolation method, but were more over-articulated than those produced by the prosodic control method. The utterances generated by the PROS and FUS+PROS methods sounded very similar, so that neither method could be preferred over the other.

The perceptual evaluation described in Section 5.3 did not include the methods based on unit fusion. This was decided to maintain a unity of approach in presenting the prosodic control method in Section 5.2, and also because the perceptual differences between the LIN, FUS, PROS, and FUS+PROS methods become quite subtle, requiring listeners to focus on specific vowels or sonorant transitions. Given the limited number of listeners available, more reliable results could be obtained if only three systems had to be compared rather than five.

In future work, we want to explore whether unit fusion can improve the performance of the prosodic control method, for example in the case where the acoustic units do not consist of diphones, but are selected from a larger corpus of fluent speech.

5.6 Discussion

We have presented a method to control the degree of articulation of sonorant units, by imposing dynamic constraints on the LSF trajectories. The method was motivated by the contextual model of articulation, according to which the degree of articulation of an acoustic unit depends on (1) the phonetic context, (2) the phoneme durations, and (3) the spectral rate-of-change. Spectral modifications were based on the numerical data obtained in Chapter 4, where the effects of prosodic factors on the spectral dynamics were analyzed for a balanced corpus.

The proposed method does not require the (artificial) segmentation of speech units into stationary and transition regions, as would be required for an approach based on non-uniform time-warping. Rather, the dynamic characteristics of human speech articulation are integrated robustly, by specifying target spectral dynamics based on the local rate-of-change of the acoustic units. The method also allows specification of spectral dynamics that *increase* the degree of articulation of acoustic units, for example to emphasize words or to adopt a clear, "didactic" articulation style. Using time-warping, no phonetic targets outside the range of the original acoustic units could be reached.

Experimental results showed that the proposed method improved the intelligibility and naturalness of concatenated speech, by controlling the degree of articulation of sonorant phonemes. Hence, the variations in the spectral dynamics measured in Chapter 4 are not merely a "production effect" of natural speech; rather, the spectral dynamics are shown to have a communicative role, by supporting the prosodic structure of a spoken message.

The techniques discussed in this chapter show that it is possible to alter the formant structure of acoustic units, for example to generate target undershoot, while preserving the perceptual quality of the original recordings. Therefore, the techniques increase the flexibility of concatenative synthesis systems, and open up new possibilities to explore the perceptual effects of quantative formant modifications in natural sounding speech.

Chapter 6

Conclusion

6.1 Summary

We have investigated how the degree of articulation of sonorant phonemes changes in natural speech, and we have proposed techniques to control the degree of articulation in computer-generated speech.

Analysis and synthesis of the degree of articulation focused on the spectral rate-of-change of acoustic units, which is related to the articulation effort used by a speaker. According to the contextual model of articulation, the degree of articulation of phonemes depends on (1) the phonetic context, (2) the phoneme durations, and (3) the spectral rate-of-change. Hence, for a given phoneme sequence and given phoneme durations, the degree of articulation can be controlled by modifying the spectral rate-of-change.

The following contributions were made in this thesis:

- A method was described to modify the spectral shape of acoustic units, while preserving a perceptual quality comparable to that of unmodified units. The method was based on warping the spectral residual using a sinusoidal + all-pole representation of speech.

- The effects of prosodic factors on the spectral rate-of-change of sonorant phoneme transitions were analyzed. The rate-of-change was shown to increase with linguistic prominence, i.e., in stressed vs. unstressed syllables, in accented vs. unaccented words, in sentence-medial vs. sentence-final words, and when a clear articulation style is used.

- A numerical model was formulated and trained to predict the rate-of-change of liquid-vowel, diphthong, and vowel-liquid transitions, depending on the prosodic context.
- The contextual model of articulation was integrated in the framework of a concatenative synthesizer, based on a line spectral frequency (LSF) representation of the acoustic units. The proposed method computes the most likely LSF trajectories given a set of static and dynamic constraints. The spectral shape of acoustic units then is modified corresponding to the target LSF trajectories.
- Concatenation mismatch between acoustic units was reduced by specifying consistent spectral dynamics at concatenation points, based on the characteristics of overlapping *fusion* units.

6.2 Future Work

In this thesis, we have analyzed and synthesized the degree of articulation of one speaker. In the experiments described in Sections 5.3 and 5.4, the acoustic units were produced by the same speaker that recorded the database in Chapter 4. Hence, the spectral dynamics predicted using Table 5.1 enabled more accurate synthesis of the articulatory behavior of this speaker. The data in Table 5.1 are also easily applied to acoustic units from different speakers. In future work, speaker-specific articulatory characteristics can be investigated. This may enable synthesis of different articulation styles, i.e., individual accents or dialects.

The spectral modification method proposed in Chapter 2 was based on a sinusoidal + all-pole representation of speech. It will be interesting to evaluate this method for high-pitched speakers, such as children, for whom the number of pitch harmonics in the speech spectrum decreases. Having a smaller number of sinusoidal parameters might complicate the estimation of the all-pole parameters (El-Jaroudi and Makhoul, 1991), and limit subsequent spectral modifications. However, in Chapter 2 an early version of the spectral modification method was tested for the male speaker used in this thesis and for a female speaker with a fairly low and breathy voice (also Wouters and Macon, 2000). The proposed method yielded similar results for both voices in a vowel transformation task.

In the present work, the spectral constraints were applied only in sonorant speech segments, as the resonant frequencies of the vocal tract cannot be modelled in the spectrum of non-sonorant segments, or must be separated from aperiodic (i.e., noise-like) signal components. Hence, the endpoints of LSF trajectories were "frozen" at the boundaries between sonorant and non-sonorant speech, often limiting the amount of spectral modification that could be achieved.

A solution to this problem would be to obtain a robust estimation of the resonant frequencies of the vocal tract, both in sonorant and non-sonorant speech, for example by measuring the geometry of the vocal tract during speech recording. Such resonance trajectories could then be associated with the high-energy peaks of the spectrum in sonorant segments, while their dynamics could be controlled across sonorant and non-sonorant speech regions, according to the contextual model of articulation. Integration of such articulatory measurements in a concatenative synthesis system is a topic for further investigation. Another possibility is to extend the proposed method to all voiced phonemes, such as /v,z,ʒ,dʒ/ in English, and the voiced /h/. This could be achieved by splitting the speech signal into a deterministic and an aperiodic component (Stylianou, 2001; Yegnanarayana *et al.*, 1998; Richard and d'Alessandro, 1996; Acero, 1998), and applying the proposed method to the deterministic component only.

In Chapter 5, several assumptions were made to extend the numerical model trained in Chapter 4 to other prosodic and phonetic contexts. These assumptions should be validated using additional data collection and analysis. However, since prosodic factors such as sentence accent and word position can produce essentially a continuous variety of prosodic contexts, design of a balanced database to study independent contributions of each prosodic factor may become unfeasible. Rather, statistical techniques to deconfound factors in a large corpus of speech, as proposed in the context of duration modeling by van Santen (1992), may be applied successfully.

Another challenge for future research is to modify additional acoustic-prosodic characteristics of speech units. For example, Sluijter and van Heuven (1996) have analyzed the effects of prosodic factors on the spectral balance, which is related to the glottal effort applied by a speaker. In Sluijter and van Heuven (1997), modifications of the spectral balance were integrated in a formant synthesis system, and were shown to improve stress perception, using an experimental setup similar to the one in Section 5.4. These modifications can be integrated in the method proposed in this thesis, and should benefit both intelligibility and naturalness.

Similarly, the ratio of the first harmonics of the fundamental frequency, which is related to the *open quotient* of the glottal waveform (Klatt and Klatt, 1990; Doval and d'Alessandro, 1997), can be easily modified in the sinusoidal + all-pole representation. Hence, acoustic-prosodic effects related to voice quality, i.e., variations from "tense" to "soft" or "breathy" speech, could be synthesized. An interesting topic for future research is to analyze the effects of prosodic factors on voice quality parameters such as the open quotient or the spectral balance, using natural speech data. This would enable control of such parameters during speech synthesis, adding further flexibility to the concatenative framework, and making collection of specialized acoustic units, such as glottalized or sentence-final units, unnecessary.

Bibliography

Acero, A., 1998, A mixed-excitation frequency domain model for time-scale pitch-scale modification of speech, *Proceedings of ICSLP*, 1923–1926.

Acero, A., 1999, Formant analysis and synthesis using Hidden Markov Models, *Proceedings of EUROSPEECH*, 1047–1050.

Allen, J., D. Klatt, and S. Hunnicutt, 1987, *From Text to Speech - The MITalk system* (Cambridge University Press).

Balestri, M., A. Pacchiotti, S. Quazza, P. L. Salza, and S. Sandri, 1999, Choose the best to modify the least: a new generation concatenative synthesis system, *Proceedings of EUROSPEECH*, 2291–2294.

Bellegarda, J. R., K. E. A. Silverman, K. Lenzo, and V. Anderson, 2001, Statistical prosodic modeling: from corpus design to parameter estimation, IEEE Transactions on Speech and Audio Processing **9**(1), 52–66.

van Bergem, D. R., 1993, Acoustic vowel reduction as a function of sentence accent, word stress and word class, Speech Communication **12**, 1–23.

Beutnagel, M., A. Conkie, and A. Syrdal, 1998, Diphone synthesis using unit selection, *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, 185–190.

Blackburn, C. S. and S. Young, 2000, A self-learning predictive model of articulator movements during speech production, Journal of the Acoustical Society of America **107**(3), 1659–1670.

Botinis, A., B. Granström, and B. Möbius, 2001, Developments and paradigms in intonation research, Speech Communication **33**(4), 263–296.

Breen, A. P. and P. Jackson, 1998a, Non-uniform unit selection and the similarity metric within BT's Laureate TTS system, *Proceedings of the Third ESCA/COCOSDA Workshop on Speech Synthesis*, 201–206.

Breen, A. P. and P. Jackson, 1998b, A phonologically motivated method of selecting non-uniform units, *Proceedings of ICSLP*, 2735–2738.

Broad, D. J. and F. Clermont, 1989, Formant estimation by linear transformation of the LPC cepstrum, Journal of the Acoustical Society of America **86**(5), 2013–2017.

Browman, C. P. and L. Goldstein, 1992, Articulatory phonology: an overview, Phonetica **49**, 155–180.

Campbell, N. and A. W. Black, 1996, *Prosody and the selection of source units for concatenative synthesis* (Springer-Verlag), chapter 22, 279–292.

Chen, J.-D. and N. Campbell, 1999, Objective distance measures for assessing concatenative speech synthesis, *Proceedings of EUROSPEECH*, 611–614.

Childers, D. G. and C. Ahn, 1995, Modeling the glottal volume-velocity waveform for three voice types, Journal of the Acoustical Society of America **97**(1), 505–518.

Cole, R., T. Carmell, P. Connors, M. Macon, J. Wouters, J. de Villiers, A. Tarachow, D. Massaro, M. Cohen, J. Beskow, J. Yang, U. Meier, *et al.*, 1998, Intelligent animated agents for interactive language training, *STiLL: ESCA Workshop on Speech Technology in Language Learning* (Stockholm, Sweden), also available at: http://cslu.ece.ogi.edu/publications [Viewed: April 2001].

Conkie, A. D. and S. Isard, 1997, Optimal coupling of diphones, *Progress in Speech Synthesis*, edited by J. P. H. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Springer-Verlag, New York), 293–304.

Delogu, C., S. Conte, and C. Sementina, 1998, Cognitive factors in the evaluation of synthetic speech, Speech Communication **24**(2), 153–168.

Donovan, R., 1998, The IBM trainable speech synthesis system, *Proceedings of ICSLP*, volume 5, 1703–1706.

Doval, B. and C. d'Alessandro, 1997, Spectral methods for voice source parameters estimation, *Proceedings of EUROSPEECH*, 533–536.

Drullman, R. and R. Collier, 1991, On the combined use of accented and unaccented diphones in speech synthesis, Journal of the Acoustical Society of America **90**(4), 1766–1775.

Dutoit, T. and H. Leich, 1994, On the ability of various speech models to smooth segment discontinuities in the context of text-to-speech synthesis by concatenation, *Proceedings of EUSIPCO*, volume 1, 8–12.

Edgington, M., 1997, Investigating the limitations of concatenative synthesis, *Proceedings of EUROSPEECH*, 593–596.

El-Jaroudi, A. and J. Makhoul, 1991, Discrete all-pole modeling, IEEE Transactions on Signal Processing **39**(2), 411–423.

Fant, G., 1960, *Acoustic Theory of Speech Production* (Mouton & Co., The Hague).

Flege, J. E., 1988, Effects of speaking rate on tongue position and velocity of movement in vowel production, Journal of the Acoustical Society of America **84**(3), 901–916.

Fourakis, M., 1991, Tempo, stress and vowel reduction in American English, Journal of the Acoustical Society of America **90**(4), 1816–1827.

Furui, S., 1986, On the role of spectral transition for speech perception, Journal of the Acoustical Society of America **80**(4), 1016–1025.

Gabioud, B., 1994, Articulatory models in speech synthesis, *Fundamentals of speech synthesis and speech recognition*, edited by E. Keller (John Wiley & Sons), chapter 10, 215–230.

Gay, T., 1968, Effect of speaking rate on diphthong formant movements, Journal of the Acoustical Society of America **44**(6), 1570–1573.

Gay, T., 1978, Effect of speaking rate on vowel formant movements, Journal of the Acoustical Society of America **63**(1), 223–230.

George, E. B. and M. J. T. Smith, 1997, Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model, IEEE Transactions on Speech and Audio Processing **5**(5), 389–406.

Gimênez, F. M., M. H. Savoji, and J. M. Pardo, 1994, New algorithm for spectral smoothing and envelope modification for LP-PSOLA synthesis, *Proceedings of ICASSP*, I.573–576.

Granström, B., 1992, The use of speech synthesis in exploring different speaking styles, Speech Communication **11**(4-5), 347–355.

Hermansky, H., 1990, Perceptual linear predictive (PLP) analysis of speech, Journal of the Acoustical Society of America **87**(4), 1738–1752.

Hermansky, H., H. Fujisaki, and Y. Sato, 1984, Spectral envelope sampling and interpolation in linear predictive analysis of speech, *Proceedings of ICASSP*, 2.2.1 – 2.2.4.

Hermansky, H. and N. Morgan, 1994, Rasta processing of speech, IEEE Transactions on speech and audio processing **2**(4), 578–589.

Hertz, S., 1991, Streams, phones and transitions: Toward a new phonological and phonetic model of formant timing, Journal of Phonetics **19**, 91–109.

Hirose, K., H. Fujisaki, and M. Yamaguchi, 1984, Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information, *Proceedings of ICASSP*, 2.13.1–2.13.4.

Hon, H., A. Acero, X. Huang, J. Liu, and M. Plumpe, 1998, Automatic generation of synthesis units for trainable text-to-speech systems, *Proceedings of ICASSP*, 293–296.

Hosom, P., 2000, *Automatic time alignment of phonemes using acoustic-phonetic information*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, 20000 Walker Rd, Beaverton, Oregon 97006 USA.

Hunt, A. J. and A. W. Black, 1996, Unit selection in a concatenative speech synthesis system using a large speech database, *Proceedings of ICASSP*, 373–376.

Itakura, F., 1975, Line spectrum representation of linear predictive coefficients of speech signals, Journal of the Acoustical Society of America **57**, S35.

ITU P.800, 1996, Methods for subjective determination of transmission quality, International Telecommunication Union (ITU), Recommendation P.800, http://www.itu.int [Viewed: April 2001].

Iwahashi, N. and Y. Sagisaka, 1995, Speech segment network approach for optimization of synthesis unit set, Computer Speech and Language **9**, 335–352.

Jilka, M., G. Möhler, and G. Dogil, 1999, Rules for the generation of ToBI-based american english intonation, Speech Communication **28**(2), 83–108.

de Jong, K. J., 1995, The supraglotal articulation of prominence in english: Linguistic stress as localized hyperarticulation, Journal of the Acoustical Society of America **97**(1), 491–504.

Klabbers, E. and R. Veldhuis, 1998, On the reduction of concatenation artefacts in diphone synthesis, *Proceedings of ICSLP*, volume 6, 2759–2762.

Klabbers, E. and R. Veldhuis, 2001, Reducing audible spectral discontinuities, IEEE Transactions on Speech and Audio Processing **9**(1), 39–51.

Klatt, D. H., 1987, Review of text-to-speech conversion for English, Journal of the Acoustical Society of America **82**(3), 737–793.

Klatt, D. H. and L. C. Klatt, 1990, Analysis, synthesis, and perception of voice quality variations among female and male talkers, Journal of the Acoustical Society of America **87**(2), 820–857.

Kohler, K. J., 1990, Segmental reduction in connected speech in German: phonological facts and phonetic explanations, *Speech production and speech modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic publishers), 69–92.

Kozhevnikov, V. A. and L. A. Chistovich, 1965, *Speech: Articulation and Perception*, JPRS 30,543 (Washington, D.C.: Joint Publication Research Service).

Ladefoged, P., 1975, *A Course in Phonetics* (Harcourt Brace Jovanovich).

Laprie, Y. and M.-O. Berger, 1996, Cooperation of regularization and speech heuristics to control automatic formant tracking, Speech Communication **19**(4), 255–269.

Laroche, J., Y. Stylianou, and E. Moulines, 1993, HNS: speech modification based on a harmonic + noise model, *Proceedings of ICASSP*, 550–553.

Lindblom, B., 1963, Spectrographic study of vowel reduction, Journal of the Acoustical Society of America **35**(11), 1773–1781.

Lindblom, B., 1990, Explaining phonetic variation: a sketch of the H&H theory, *Speech Production and Speech Modelling*, edited by W. Hardcastle and A. Marchal (Kluwer Academic Publishers), 403–439.

Macon, M. W., 1996, *Speech Synthesis Based on Sinusoidal Modeling*, Ph.D. thesis, Georgia Institute of Technology.

Makhoul, J., 1975, Linear prediction: A tutorial review, Proceedings of the IEEE **63**(4), 561–580.

McAulay, R. and T. F. Quatieri, 1995, Sinusoidal coding, *Speech Coding and Synthesis*, edited by W. Kleijn and K. Paliwal (Elsevier), 121–173.

McAulay, R. J., 1984, Maximum likelihood spectral estimation and its application to narrow-band speech coding, IEEE Transactions on Acoustics, Speech, and Signal Processing **34**(4), 744-754.

McAulay, R. J. and T. Quatieri, 1986, Speech analysis/synthesis based on a sinusoidal representation, IEEE Transactions on Acoustics, Speech, and Signal Processing **34**(4), 744-754.

Moon, S.-J. and B. Lindblom, 1994, Interaction between duration, context, and speaking style in english stressed vowels, Journal of the Acoustical Society of America **96**(1), 40–55.

Moulines, E. and F. Charpentier, 1990, Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones, Speech Communication **9**(5/6), 453–467.

Murthy, H. A. and B. Yegnanarayana, 1991, Formant extraction from group delay function, Speech Communication **10**(3), 209-221.

Nakajima, S., 1994, Automatic synthesis unit generation for English speech synthesis based on multi-layered context oriented clustering, Speech Communication **14**, 313–324.

Nakajima, S. and H. Hamada, 1988, Automatic generation of synthesis units based on context oriented clustering, *Proceedings of ICASSP*, 659–662.

Nelson, W. L., 1983, Physical principles for economies of skilled movements, Biological Cybernetics **46**, 135–147.

Nooteboom, S. G., 1997, Section introduction. Text and prosody, *Progress in speech synthesis*, edited by J. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg (Springer-Verlag, New York), chapter 6, 431–434.

Paliwal, K. K., 1995, Interpolation properties of linear prediction parametric representations, *Proceedings of EUROSPEECH*, 1029–1032.

Perrier, P. and D. J. Ostry, 1994, Dynamic modelling and control of speech articulators: application to vowel reduction, *Fundamentals of speech synthesis and speech recognition*, edited by E. Keller (John Wiley & Sons), chapter 11, 231–251.

Pierrehumbert, J., 1981, Synthesizing intonation, Journal of the Acoustical Society of America **70**(4), 985–995.

Pitermann, M., 2000, Effect of speaking rate and contrastive stress on formant dynamics and vowel perception, Journal of the Acoustical Society of America **107**(6), 3425–3437.

Plumpe, M., A. Acero, H. Hon, and X. Huang, 1998, HMM-based smoothing for concatenative speech synthesis, *Proceedings of ICSLP*, 2751–2754.

Portele, T. and B. Heuft, 1997, Toward a prominence-based synthesis system, Speech Communication **21**, 61–72.

Quatieri, T. F. and R. J. McAulay, 1987, Mixed-phase deconvolution of speech based on a sine-wave model, *Proceedings of ICASSP*, 649–652.

Quatieri, T. F. and R. J. McAulay, 1992, Shape invariant time-scale and pitch modification of speech, IEEE Transactions on Signal Processing **40**(3), 497–510.

Richard, G. and C. d'Alessandro, 1996, Analysis/synthesis and modification of the speech aperiodic component, Speech Communication **19**, 221–244.

Ross, K., 1994, *Modeling of Intonation for Speech Synthesis*, Ph.D. thesis, Boston University, College of Engineering.

van Santen, J. P. H., 1992, Contextual effects on vowel duration, Speech Communication **11**(6), 513–546.

van Santen, J. P. H., 1993, Perceptual experiments for diagnostic testing of text-to-speech systems, Computer Speech and Language **7**(1), 49–100.

van Santen, J. P. H., 1997, Prosodic modeling in text-to-speech synthesis, *Proceedings of EUROSPEECH*, Keynote 19–28.

Scha, R., 1992, Virtual voices, Mediamatic **7**(1), also online at: http://www.hum.uva.nl/computerlinguistiek/scha/IAAA/virtual.html [Viewed: April 2001].

Schulman, R., 1988, Articulatory dynamics of loud and normal speech, Journal of the Acoustical Society of America **85**(1), 295–312.

Silverman, H. and D. Morgan, 1990, The application of dynamic programming to connected speech recognition, IEEE ASSP Magazine **7**(3), 6–25.

Silverman, K., S. Basson, and S. Levas, 1990, Evaluating synthesiser performance: is segmental intelligibility enough?, *Proceedings of ICASSP*, 981–984.

Sluijter, A. and V. van Heuven, 1996, Spectral balance as an acoustic correlate of linguistic stress, Journal of the Acoustical Society of America **100**(4), 2471–2485.

Sluijter, A. and V. van Heuven, 1997, Spectral balance as a cue in the perception of linguistic stress, Journal of the Acoustical Society of America **101**(1), 503–513.

van Son, R. and L. Pols, 1990, Formant frequencies of Dutch vowels in a text, read at normal and fast rate, Journal of the Acoustical Society of America **88**, 1683–1693.

van Son, R. and L. Pols, 1992, Formant movements of Dutch vowels in a text, read at normal and fast rate, Journal of the Acoustical Society of America **92**, 121-127.

van Son, R. and L. Pols, 1999, An acoustic description of consonant reduction, Speech Communication **28**(2), 125–140.

Soong, F. K. and B.-H. Juang, 1984, Line spectrum pairs (LSP) and speech data compression, *Proceedings of ICASSP*, 1.10.1–1.10.4.

Sproat, R., M. Ostendorf, and A. Hunt, editors, 1999, *The need for increased speech synthesis research* (report of the 1998 NSF workshop for discussing research priorities and evaluation strategies in speech synthesis), available at: http://cslu.ece.ogi.edu/publications [Viewed: June 2001].

Sproat R., editor, 1998, *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (Kluwer Academic Publishers).

Strange, W., 1989a, Dynamic specification of coarticulated vowels spoken in sentence context, Journal of the Acoustical Society of America **85**(5), 2135–2153.

Strange, W., 1989b, Evolving theories of vowel perception, Journal of the Acoustical Society of America **85**(5), 2081–2087.

Stylianou, Y., 1996, *Harmonic Plus Noise Models for Speech, Combined with Statistical Methods for Speech and Speaker Modification*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, France.

Stylianou, Y., 1999, Assessment and correction of voice quality variabilities in large speech databases for concatenative speech synthesis, *Proceedings of ICASSP*, 377–380.

Stylianou, Y., 2001, Applying the harmonic plus noise model in concatenative speech synthesis, IEEE Transactions on Speech and Audio Processing **9**(1), 21–29.

Stylianou, Y., O. Cappé, and E. Moulines, 1995a, Statistical methods for voice quality transformation, Proceedings of EUROSPEECH , 447–450.

Stylianou, Y., J. Laroche, and E. Moulines, 1995b, High quality speech modification based on a harmonic + noise model, Proceedings of EUROSPEECH , 451–454.

Syrdal, A., Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, 1998a, TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis, *Proceedings of ICASSP*, 273–276.

Syrdal, A. K., A. Conkie, and Y. Stylianou, 1998b, Exploration of acoustic correlates in speaker selection for concatenative synthesis, *Proceedings of ICSLP*, 2743–2746.

Syrdal, A. K., C. W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. J. Makashay, 2000, Corpus-based techniques in the AT&T nextget synthesis system, *Proceedings of ICSLP*, paper nr. 3601.

Takano, S. and M. Abe, 1999, A new F0 modification algorithm by manipulating harmonics of magnitude spectrum, *Proceedings of EUROSPEECH*, 1875–1878.

Takano, S., K. Tanaka, H. Mizuno, M. Abe, and S. Nakajima, 2001, A Japanese TTS system based on multiform units and a speech modification algorithm with harmonics reconstruction, IEEE Transactions on Speech and Audio Processing **9**(1), 3–10.

Takeda, K., K. Abe, and Y. Sagisaka, 1992, On the basic scheme and algorithms in non-uniform unit speech synthesis, *Talking machines: theories, models, and designs*, edited by T. S. G. Bailly, C. Benoit (Elsevier), 93–105.

Talkin, D., 1995, A robust algorithm for pitch tracking (RAPT), *Speech coding and synthesis* (Elsevier), 495–518.

Traunmüller, H., 1994, Conventional, biological and environmental factors in speech communication: a modulation theory, Phonetica **51**, 170–183.

Traunmüller, H. and A. Eriksson, 2000, Acoustic effects of variation in vocal effort by men, women, and children, Journal of the Acoustical Society of America **107**(6), 3438–3451.

Tseng, C., 1995, A phonetically oriented speech database for Mandarin Chinese, *Proceedings of the International Congress of Phonetic Sciences*, 326–329.

Waibel, A., 1986, Recognition of lexical stress in a continuous speech understanding system – a pattern recognition approach, *Proceedings of ICASSP*, 2287–2290.

Watson, C. I. and J. Harrington, 1999, Acoustic evidence for dynamic formant trajectories in Australian English vowels acoustic evidence for dynamic formant trajectories in australian english vowels, Journal of the Acoustical Society of America **106**(1), 458–468.

Welling, L. and H. Ney, 1998, Formant estimation for speech recognition, IEEE Transactions on Speech and Audio Processing **6**(1), 36–48.

Widera, C., 2000, Strategies of vowel reduction – a speaker-dependent phenomenon, *Proceedings of ICSLP*, I.552–555.

Wouters, J. and M. W. Macon, 1998, A perceptual evaluation of distance measures for concatenative speech synthesis, *Proceedings of ICSLP*, volume 6, 2747–2750.

Wouters, J. and M. W. Macon, 2000, Spectral modification for concatenative speech synthesis, *Proceedings of ICASSP*, II.941–944.

Wouters, J. and M. W. Macon, 2001a, Control of spectral dynamics in concatenative speech synthesis, IEEE Transactions on Speech and Audio Processing **9**(1), 30–38.

Wouters, J. and M. W. Macon, 2001c, Effect of prosodic factors on spectral dynamics. Part I: Analysis, Journal of the Acoustical Society of America (submitted).

Wouters, J. and M. W. Macon, 2001b, Effect of prosodic factors on spectral dynamics. Part II: Synthesis, Journal of the Acoustical Society of America (submitted).

Wrede, B., G. A. Fink, and G. Sagerer, 2000, Influence of duration on static and dynamic properties of German vowels in spontaneous speech, *Proceedings of ICSLP*, I.82–85.

Yegnanarayana, B., C. d'Alessandro, and V. Darsinos, 1998, An iterative algorithm for decomposition of speech signals into periodic and aperiodic components, IEEE Transactions on Speech and Audio Processing **6**(1), 1–11.

Yong, M., 1994, A new LPC interpolation technique for CELP coders, IEEE Transactions on Communications **42**(1), 34–38.

Biographical Note

Johan Wouters was born in Arolsen, Germany, on June 6, 1972. He received the master of science degree in Electrical Engineering from the Katholieke Universiteit Leuven (KUL), Belgium, in 1996, graduating *magna cum laude.* His research interests include speech synthesis, digital signal processing, speech recognition, models of human speech production and perception, and human computer interfaces. He is a member of the International Speech Communication Association (ISCA) and of the Institute of Electrical and Electronics Engineers (IEEE). At the Oregon Graduate Institute, Johan has worked on the development of several synthesis voices in American English, Mexican Spanish, and Brazilian Portuguese. He has also worked on TTS mark-up, extending the Sable mark-up standard, implementing a module in the Festival speech synthesis system to interpret the new commands, and designing a graphical user interface to support creation and visualization of marked-up text documents. In the summer of 1998, Johan was a summer intern with Sun Microsystems, working on the Java Speech Application Program Interface (JSAPI). In April 2000, he visited Chile to collaborate with colleagues at the University of Playa Ancha in Valparaiso, to apply spoken language technology in classrooms with profoundly deaf children. He also visited the SPOLTECH group at the Universidade Federal do Rio Grande do Sul to develop a Brazilean Portuguese TTS system. Johan is a first author on one published journal article and two articles under revision, as well as five conference publications. He is a coauthor on four other conference papers and one U.S. Patent.