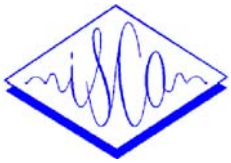

5th ISCA Speech Synthesis Workshop
Pittsburgh, PA, USA
June 14-16, 2004

IMPROVING TTS BY HIGHER AGREEMENT BETWEEN PREDICTED VERSUS OBSERVED PRONUNCIATIONS

Yeon-Jun Kim, Ann Syrdal

AT&T Labs-Research,
180 Park Ave. Florham Park, NJ 07932

Matthias Jilka

Institut für Linguistik, Universität Stuttgart,
Keplerstraße 17, Stuttgart, Germany

ABSTRACT

This paper looks at improving unit selection text-to-speech (TTS) quality by optimizing the agreement between front-end and speech database.

We focused, in particular, on two classes of problems causing degradation in synthesis quality: 1) realization of /d/ and /t/[1] sounds and 2) confusions of unstressed vowels, especially with *schwas*.

We investigated two approaches to tackling these problems. First, we improved the phonological processing in the front end modules. Further improvement resulted from creating *speaker-dependent* pronunciation lexicons for automatic speech labeling of our voice databases. This change helped in alleviating many pronunciation errors that resulted from mismatches between *lexical* pronunciations and how the speaker (voice talent) *actually* pronounced a word, while keeping consistency in labeling. Each speaker has his or her own unique pronunciations (and context-dependent variations), so that no one standard lexicon is able to cover all of the speakers' variations.

A subjective listening test showed that combining these two approaches resulted in perceived quality improvement for American English male and female voices.

1. INTRODUCTION

Since the introduction of unit selection concatenative synthesis techniques into TTS research, the synthesis quality of TTS systems has improved dramatically. Currently, many commercial TTS systems employ unit selection synthesis techniques and deliver high quality synthetic speech [1].

Even though TTS systems have improved and are commercially used, the output synthesis quality of current TTS systems is still far from natural human speech [2]. To improve synthesis quality of a unit selection TTS system, we have analyzed synthesis data and found a number of general problems. The analysis reveals problems that fall into two broad categories: First, problems related with unnatural prosody, such as lack of prominence and the same prosody patterns in every sentence, are not new. To make the TTS system synthesize prosodically natural speech, many techniques, such as prosody modification and voice rendering, have been studied for a long time and brought some level of successes, but there are still unsolved problems remaining.

The other type of problem is related to contextual mismatches that a unit asked by a TTS front-end is contextually different from a unit chosen by a synthesis module. It has been known as a typical problem of diphone-based approach and resolved by introducing of unit selection approach. Unit selection approach chooses the best unit considering phonetic- and prosodic-context most of the time, and synthesizes very natural speech when it chooses contiguous units. However, a unit selection TTS system sometimes produces conspicuous errors which may be less noticeable in low quality diphone-based synthesis.

Most of unit selection TTS errors are caused by mismatches between pronunciation prediction and database labels. Units in the large-scale speech inventory can be labeled differently from a speaker's actual pronunciations, whereas each unit in a small TTS speech inventory was very carefully labeled and agreed with its front-end. Higher agreement between pronunciation module and labels in the speech inventory is critical to reduce the number of mistakes in a unit selection TTS system as described with regards to *German* in [3].

This paper describes our efforts to reduce mismatches between predicted pronunciation and speech inventory labels in two *American English* voices leading to improved synthesis quality in our unit selection TTS.

In section 2, the requirements for reducing mismatches in unit selection TTS synthesis are described. In order to detect these mismatches in large-scale speech corpora, pronunciation variants suggested by automatic speech recognition (ASR) are analyzed in section 3. Section 4 deals with acoustic model training for our particular TTS, which is quite different from that for "traditional" ASR. This paper concludes with a perceptual evaluation of synthetic quality based on the voice databases created using the methods described.

[1] In this paper, we use the DARPABET symbols for specifying phones which include allophonic variations.

2. HIGHER AGREEMENT OF FRONT-END AND LABELS IN VOICE INVENTORIES

In bridging the gap between phonological descriptions predicted by the TTS front-end and phone labels of actual speech, there are three aspects that need to be controlled: 1) the TTS letter-to-sound conversion, 2) the phonetic transcriptions provided for automatic labeling, 3) the speaker's actual productions [3].

2.1. Detailed Phone-Set for Letter-to-Sound

If we use rule-based (possibly statistical) letter-to-sound for the purpose of transcriptions of formal pronunciations found in the standard dictionary, pronunciation discrepancies between front-end prediction and speech inventories may be unavoidable [4].

Two particular patterns emerged as problems in earlier front-ends which often transcribed with phonemic (abstract) symbols rather than phonetic ones (detailed allophonic variants); 1) realization of /d/ and /t/ sounds and 2) confusions of unstressed vowels. especially with *schwas*.

The phone /t/ in American English has three distinctive phonetic realizations in the 'detailed phone-set': "normal" /t/ (with stop closure and release), flapped [dx], glottalized [q]. The flapped [dx] is an allophonic variation of /t/ in the word 'butter'. The other variation would be the glottalized [q] such as in the middle of the word 'cotton'. When [q] happens to be chosen by unit-selection where normal /t/ should be used, a listener will perceive that as a missing /t/.

The reduced fronted vowel, [ix], only in the 'detailed phone-set', was another source of confusion. An example of allophone [ix] is the vowel in the final syllable of the word "abandon" (vs. the non-fronted /ax/ schwa in the beginning of the same word). For example, /uw/ in "contributed" usually reduces to [ix] in casual speech even though its lexical transcription is /k ax n 0 t r ih 1 b y uw 0 t ih d 0/. When unit-selection chose /b y uw/ in "contributed" for synthesizing "beautiful", it would sound like /b y ix dx ih f ax l/.

In earlier TTS versions, our front-end transcribed with *phonemic* symbols that did not include allophonic variations. We also labeled speech databases with phonemic symbols and relied on phonetic context and unit-selection to choose the correct phone for a given context. Therefore, there was degradation in synthesis quality when unit-selection failed to choose the correct phonetic realizations. This problem even affected intelligibility in some cases, not only degraded voice quality.

Allophones [dx], [q], [ix] in the 'detailed phone-set' have been added in the new TTS system so that unit-selection should choose the correct sounds as defined by the front-end. The post-lexical phonology module has also been updated in line with the newly introduced allophones.

Besides these allophones we added, there are several more allophones in the DARPABET such as [nx], [hw], [ux], etc., but we consider only [dx], [q], [ix] here.

2.2. Pronunciation Lexicon for Automatic Labeling

A pronunciation lexicon is one of the core components in automatic phone labeling of speech. Its purpose is to map the orthographic representation of a word to its pronunciation. Automatic phone labeling for speech synthesis is commonly done by *forced alignment* [6], so that each phone label is determined not only by acoustic models but also by a transcription in the pronunciation lexicon even if the transcription does not match with the speech signal.

The requirements of a pronunciation lexicon for automatic labeling are as follows:

- A transcription in the lexicon for automatic labeling needs to be as close as possible to the one predicted by the TTS front-end so that TTS (unit-selection) can choose the best units for synthesizing a word.
- A transcription should also accurately describe speakers' pronunciations for higher agreement between front-end and speech database.

Based on the requirements described above, the previous pronunciation lexicon that consisted of transcriptions used in the TTS front-end and an ASR lexicon having pronunciation variants.

The ASR lexicon contains numerous pronunciation variations generated with decision tree based phoneme-to-phone mappings [7]. It turns out, however, that the overly generated (implied) variants, which the speaker never actually produced, only complicate the correct recognition of all other variations that the speaker actually produced. Because it is difficult to focus on single speaker-specific pronunciation variants, it may not be advisable to label a TTS voice database and train the speaker-dependent acoustic models with a large multiple speakers' transcription system such as the one used for the TIMIT corpus.

To alleviate the over-generation, the new pronunciation lexicon for automatic labeling consists of 1) transcriptions from TTS pronunciation modules, 2) the PRONLEX[2] to cover well-known pronunciation variations [8], and 3) the target speaker's own pronunciation variants.

To achieve consistent transcriptions in line with a 'detailed phone' set, we apply English allophone rules [9] to PRONLEX, which is transcribed with phonemic symbols.

It is well-known in ASR that lexicon tuning concerning pronunciation variations is a tedious and labor-intensive part of building a lexicon [10]. To save time and effort, we utilize

[2]The PRONLEX from LDC is one of the best speech recognition lexicon which has about 90,000 entries.

ASR techniques for the detection of the target speaker's own pronunciation variants described in more detail in section 3.

2.3. Speaker's Actual Productions

Apart from careful monitoring of the recording sessions, the only practical way to achieve accurate transcriptions is to detect speaker-specific pronunciation variants in recorded speech using ASR techniques and add those variants into the pronunciation lexicon described in 2.2. In this paper, we describe only the part after recordings have been finished.

3. ANALYSIS OF SPEAKER-SPECIFIC PRONUNCIATION VARIANTS

For a unit selection TTS system having several hundred thousand labeled units in each speaker's voice inventory, it is impractical to check transcription variants without ASR techniques. In this work, we use an HMM-based phone recognizer to detect speaker-specific pronunciation variants. The HMMs used in this work are commonly defined as three-state left-to-right models with multiple mixture Gaussian densities. We use the 13 standard HMM input parameters: 12 MFCCs (Mel frequency cepstral coefficients) plus normalized energy, and these parameters' first and second order delta coefficients.

The variants suggested by a phone recognizer can be considered unpredictable in the sense that they are determined by the speaker's performance. Speaker performance is of course influenced by such various factors as discourse situation, speaking style or even level of concentration. Therefore, the analysis of variants produced by the phone recognizer is more experimental rather than based on theory-oriented approaches such as the explicit distinction of lexical and post-lexical variation.

3.1. Identifying pronunciation variants

Here we consider the case when the phone recognizer has correctly detected a pronunciation variant produced by the speaker that is not covered by the pronunciation lexicon. While it is certain that these free variants will occur in spoken language due to the influence of discourse situation or speaking style, it cannot be predicted when, where, or how often.

The examples below occur in our female voice database.

- *function words* are reduced in multiple ways, e.g. "and" as /eh n d/, /eh n/, /en/, or "because" as /b ax 0 k ao z 1/, /b ix 0 k ah z 1/
- *vowel reduction* generally creates a lot of variation, e.g. "political" (/p aa 0 l ih 1 t ih 0 k ax l 0/, /p ax 0 l ih 1 dx ih 0 k ax l 0/, /p ax 0 l ih 1 t ax 0 k ax l 0/), "prudential" (/p r uh 0 d eh n 1 sh ax l 0/, /p r ax 0 d eh n 1 sh ax l 1/), especially in the case of schwa alternatives or fronted schwa [ix] and unstressed /ih/, e.g., "biggest" as /b ih 1 g ih s t 0/ or /b ih 1 g ix s t 0/, "deliver" as /d ax 0 l ih 1 v er 0/ or /d ih 0 l ih 1 v er 0/
- *schwa-elisions before /r/*, e.g., "century" as /s eh n 1 ch r iy 0/, "commercially" as /k ax 0 m er sh 1 l iy 0/, "dangerous" as /d ey n jh 1 r ax s 0/, "delivery" as /d ax 0 l ih v 1 r iy 0/, "summary" as /s ah m 1 r iy 0/
- *cluster simplifications*, e.g., "impacts" as /ih m 1 p ae k s 2/, "government" as /g ah 1 v er 0 m ih n t 0/, "products" as /p r aa 1 d ax k s 0/, especially in the case of two alveolar sounds at the end of a word, "around" as /ax 0 r aw n 1/, "first" as /f er s 1/, "analysts" as /ae 1 n ax 0 l ix s 0/, "journalists" as /jh er 1 n ax 0 l ix s 0/
- varying (style-dependent) degrees of *combinatory phenomena* between words, especially *assimilations* ("would you" as /w uh jh uw/) but also reductions ("do you" as /d y uw/)

3.2. Identifying speaker's mistakes

We define speaker errors to be genuine reading errors by the speaker that went uncorrected during the recording. Such errors and inconsistent pronunciations in general should be avoided by means of strict monitoring of the recordings, but, of course, it is not possible to achieve that completely.

These kinds of errors are of course not only predictable, but it can be stated that they are more likely in unfamiliar words, like proper names, foreign words and technical vocabulary in general. In such cases the speaker can even produce more than one variation due to his/her uncertainty about the correct pronunciation.

The examples of inconsistent pronunciations given here were provided by a female speaker of American English.

- "Alvarez" as /aa l 1 v ax 0 r eh z 0/, /ae l 0 v ae 1 r eh z 0/, /ao l 1 v ax 0 r eh z 0/,
- "Bantu" as /b aa n 1 t ih 0/, /b ae n 1 t uh 0/,
- "Gendarme" as /zh ae n 1 d aa r m 2/, /zh eh n 2 d aa r m 1/,
- "Uppsala" as /uh p 1 s ax 0 l ax 0/, /ah p 2 s ey 1 l ax 0/

There is also the possibility of careless reading, a type of error that is difficult, if not impossible, to predict, e.g.

- "Fibreboard" as "Fireboard", "veterinarian" as "vegetarian"

3.3. Identifying recognition errors

The recognizer routinely mis-transcribes certain pronunciation variations that the speaker never actually produced. They mostly involve misinterpretations of transitions between phones.

These kinds of pronunciation variants suggested by the phone recognizer are thus disregarded in the continuing recognition process. Many of these errors were automatically identified by the dynamic programming (DP) matching between forced alignment result and phone recognizer transcription, but some errors needed to be corrected by hand.

- *insertion of a intrusive plosive* between a nasal and an fricative (especially /s/ or /th/), e.g., “chance” as /ch aa n **t** s/, “license” as /l ay s ax n **t** s/, “concerned” as /k ax n **t** s er n d/, “means” as /m iy n **d** z/, “amongst” as /ax m aa ng **k** s t/, “strength” as /s t r eh ng **k** th/

- loss of an *unreleased final plosive* before an initial plosive, nasal, or dental fricative, e.g., “assigned task” as “assign task”, “forced to go” as “force to go”, “returned there” as “return there”, “send mail” as “sen mail”

- plosive and fricative separated by a *morpheme boundary* interpreted as an affricate, e.g., “friendship” as /f r eh n 1 **ch** ih p 2/, “roadshow” as /r ow 1 **ch** ow 2/

- *doubled intervocalic consonant* when the following syllable has primary stress, e.g., “initial” as /ih n 0 n ih 1 sh ax l 0/, “immense” as /ih m 0 m eh n s 1/, “amount” as /ax m 0 m aw n t 1/

Other phonologically predictable cases are:

- an unaspirated voiceless plosive following /s/ in a syllable onset is *interpreted as voiced*, e.g., “scales” as /s g ey l s/, “scrambled” as /s g r ae m b ax l d/, “space” as /s b ey s/, “stood” as /s d uh d/

- a strongly aspirated /t/ is transcribed as /th/, e.g., “tan” as /th ae n/, “target” as /th aa r g ih t/, “too” as /th uw/

- stressed “dr” misinterpreted as “tr”, due to retroflex /r/, e.g., “draw” as /**t** r ao/, “drink” as /**t** r ih ng k/, “drive” as /**t** r ay v/

Linguistic knowledge and analysis of recognition results might allow the development of rules to automatically identify certain combinations of phonemes in transcriptions that either should be disregarded in the recognition process or be accepted as highly probable variants.

4. ACOUSTIC MODEL ADAPTATION WITH SPEAKER-SPECIFIC PRONUNCIATION

Speaker-specific pronunciation variants aren’t only added to the lexicon for labeling, but also help to tune speaker-dependent acoustic models. Though the number of wrong labels in initial model training is small with respect to the total number of correct phone labels, mislabels in training contaminate acoustic models. We have seen improvement in phone labeling accuracy after acoustic training with the speaker-specific pronunciation lexicon. The procedure for adapting an acoustic model is described below.

As mentioned in section 3, we use an HMM-based phone recognizer to detect speaker-specific pronunciation variants. The phone recognition procedure consists of two phases: 1) Acoustic (HMM) phone modeling and 2) Phone recognition.

Initially, only existing speaker-independent (SI) HMMs and a basic pronunciation lexicon derived from a letter-to-sound module are used to produce seed phone labels by means of Viterbi alignment for speaker-dependent (SD) HMM training. The resulting speaker-dependent HMMs are trained to provide the segmentation for building an inventory of synthesis units. During the initial training a false mapping between a word and the phonetic units may contaminate the HMMs. But, the SD HMMs are eventually fine-tuned to produce optimal results by using the phone labels from the previous iteration as the input for HMM initialization and re-estimation [11].

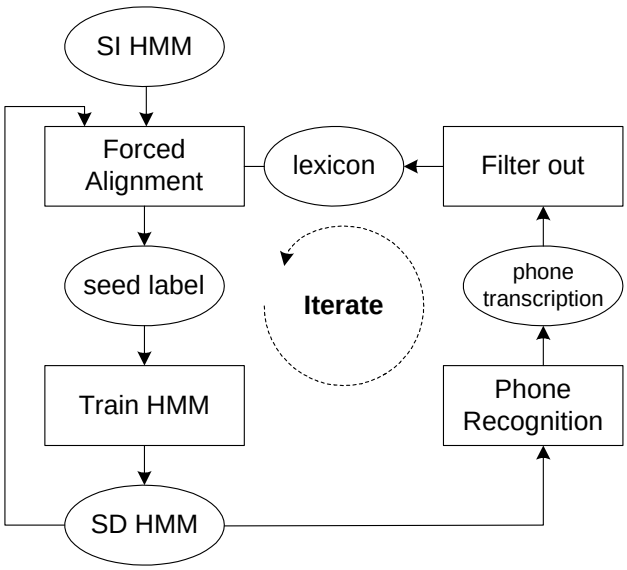

Fig. 1. Iterative acoustic model adaptation

In a second phase, phone recognition is performed with the trained SD HMMs and a phone bi-gram language model (LM) as described in [12]. A phone recognizer with a phone

bi-gram LM as opposed to *forced* alignment with a given lexicon can increase the degrees of freedom in the pronunciation transcription. However, differences between actual pronunciations and the dictionary pronunciations of words can sometimes be compensated for by using multiple Gaussian distributions in context-dependent phone HMM models.

The pronunciation variants described in section 3 serve as feedback to the pronunciation lexicon used in the next HMM training phase illustrated in Fig. 1.

As a result of the analysis, the pronunciation lexicon is improved and the process is repeated, yielding a more accurate acoustic model. Iterations are continued until there is no more improvement in the synthetic speech or no more acceptable variations are produced. These two stop conditions can be interpreted as a measure of the effectiveness of the process.

In summary, our objective is best achieved by an iterative process involving three main elements described above: (1) acoustic model training (using Hidden Markov Models (HMM's)) with a high-quality dictionary and letter-to-sound rules that cover all phonologically predictable variants of a word, (2) phone recognition with the speaker-dependent HMM's, and then (3) analysis of the recognized pronunciation variants not covered by the given pronunciation lexicon.

5. PERCEPTUAL EVALUATION

A perceptual experiment was performed to confirm that our work to achieve higher agreement between the TTS front-end and the speaker's production actually lead to a measurable improvement in synthesis quality.

The experiment elicited listener ratings for three versions each of male and female American English synthesized speech. The three versions represents as follows:
1) the reference front-end + the reference voice database
2) the reference front-end + the modified voice database
3) the modified front-end + the modified voice database

where the modified front-end transcribed with 'detailed phones' instead of phonemic symbols, and the modified voice database means the database was re-labeled with 'detailed phones' and the target speaker-specific lexicon.

5.1. Test Materials

Twelve test sentences were synthesized per voice(2) and TTS(3) version, resulting in a total of 72 test trials. Eight of the utterances were 18- to 32-word sentences selected from recent news articles covering a variety of topics (politics, international news, science, and entertainment). Two test utterances were 22- to 24-word sentences selected from Aesop's fables in the IPA manual.

The other two test utterances, from 17- to 24- words in length and composed of one or two sentences, were styled after interactive service announcements. They included explanations and questions in a speech dialog.

5.2. Listeners and Test Procedures

Eighteen speakers of English voluntarily participated in the test; ten were native speakers of English, and eight were non-native speakers. All were fluent in English, and most were also unfamiliar with the TTS system and the voices tested. A total of 1296 judgments were collected from the eighteen subjects.

The interactive and self-paced listening test was web-based, and generally lasted from 15 to 20 minutes. Written instructions described the test procedure to subjects. Listeners were directed to listen to test utterances (in any order and as many times as they liked) by clicking on sound icons.

Finally, subjects were instructed to rate the speech quality of an utterance on a 5-point rating scale (1 = bad, 2 = poor, 3 = fair, 4 = good, and 5 = excellent) by clicking on the associated radio button.

5.3. Subjective Listening Test Results

The results presented reflect the entire group of eighteen listeners. Listeners' ratings were analyzed with a fully factorial repeated measures ANOVA, in which Voice (2), TTS Version (3), and Sentences (12) were repeated factors. Two of the three main effects, TTS Version ($F(2,34) = 22.977$, $p < 0.0001$), and Sentences ($F(11,187) = 10.208$, $p < 0.0001$), were statistically significant.

There was no significant difference between overall ratings of the male speaker and the female speaker. Post hoc tests indicated that the three TTS systems differed significantly from each other: version 3 (mean rating = 3.58) was significantly better than version 1 (3.33) and 2 (3.20), and version 1 was significantly better than version 2. The Sentences main effect indicated that rated quality of the 12 test sentences differed significantly among each other.

There was a significant Voice by TTS Version interaction ($F(2,34) = 17.084$, $p < 0.0001$), reflecting the fact that 1) for TTS version 3, the male voice was rated significantly higher than the female voice, 2) the female version 3 was rated significantly higher that the female version 2, and 3) the male version 3 was significantly higher than versions 1 and 2, and the male version 1 was significantly better than version 2. A significant Voice by Sentence interaction ($F(11,187) = 3.218$, $p < 0.0001$), indicated that there were significant differences in ratings of sentences between male and female voices.

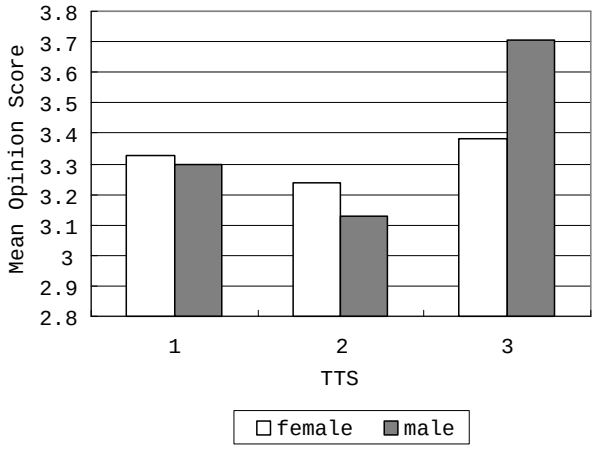

Fig. 2. Comparison of TTS systems with the male and the female voices: 1) the reference front-end + the reference voice database, 2) the reference front-end + the modified voice database, 3) the modified front-end + the modified voice database

6. SUMMARY AND CONCLUSIONS

In this paper, we described methods for achieving a higher agreement between front-end module and labels in voice databases in order to reduce mistakes in a unit selection TTS system. Efforts to reduce mismatches between improved front-end prediction and speech inventory labels in the American English voice building process led to improved synthetic quality in our unit selection TTS. To reduce mismatches, we focused, in particular, on two classes of problems; 1) realization of /d/ and /t/ sounds and 2) confusions of unstressed vowels, especially with *schwas*.

This paper also presents a procedure for recognizing speakers' pronunciation variations in order to build more accurate voice databases for unit selection TTS synthesis. The main principle described is an iterative training of acoustic models and correction of pronunciation variants. The approach not only provides higher-quality synthesis, but also constitutes a method that does not demand extensive amounts of time and effort, thus facilitating the relatively swift production of new voices.

Special emphasis was put on linguistic rules to effectively detect particularly frequent mismatches between recognition dictionary and actual speech. This may allow more time to search for less predictable deviations.

In an subjective test listening, synthetic quality was rated higher when there was a higher agreement between TTS front-end and the speaker's actual production, and particularly when a more detailed phone set was used.

7. REFERENCES

[1] Ann K. Syrdal, Colin W. Wightman, Alistair Conkie, Yannis Stylianou, Mark Beutnagel, Juergen Schroeter, Volker Strom, and Ki-Seung Lee, "Corpus-based Techniques in the AT&T NEXTGEN Synthesis System," in *Proc. ICSLP 2000, Beijing*, 2000.

[2] Juergen Schroeter, Alistair Conkie, Ann Syrdal, Mark Beutnagel, Matthias Jilka, Volker Strom, Yeon-Jun Kim, Hong-Goo Kang, and David Kapilow, "A Perspective on the Next Challenges for TTS Research," in *IEEE 2002 Workshop on Speech Synthesis*, 2002.

[3] Matthias Jilka and Ann Syrdal, "The AT&T German Text-to-Speech System: Realistic Linguistic Description," in *Proceedings of ICSLP*. 2002, Denver.

[4] Corey Andrew Miller, *Pronunciation Modeling in Speech Synthesis*, Ph.D. thesis, Univ. Pennsylvania, 1998.

[5] W. Fisher, V. Zue, D. Bernstein, and D. Pallet, "An Acoustic-Phonetic Database," *J. Acoust. Soc. Am.*, , no. 81, 1987.

[6] John Kominek, Christina Bennett, and Alan W. Black, "Evaluating and Correcting Phoneme Segmentation for Unit Selection Synthesis," in *Proceedings ESCA Eurospeech'03*, 2003.

[7] M.D. Riley and A. Ljojle, *Automatic generation of detailed pronunciation lexicons*, chapter 12, Kluwer Academic Publishers, 1995.

[8] Linguistic Data Consortium, *COMLEX English Pronuncing Lexicon*, Trustees of the University of Pennsylvania, version 0.3, 1997.

[9] P. Ladefoged, *A Course in Phonetics*, New York: Harcourt, Brace, and Jovanovich, 1993.

[10] K. L. Markey and W. Ward, "Lexical Tuning based on Triphone Confident Estimation," in *Proceedings ESCA Eurospeech'97*, 1997.

[11] Yeon-Jun Kim and Alistair Conkie, "Automatic Segmentation combining an HMM-based Approach and Spectral Boundary Correction," in *Proceedings of ICSLP*. 2002, Denver.

[12] M. Ravishankar and M. Eskenazi, "Automatic Generation of Context-dependent Pronunciation," in *Proceedings ESCA Eurospeech'97*, 1997.