Developing a Test Bed of English Text-to-Speech System XIMERA for the Blizzard Challenge 2006

Tomoki Toda[†,‡,‡‡], *Hisashi Kawai*[‡,‡‡], *Toshio Hirai*[‡‡‡], *Jinfu Ni*[†,‡], *Nobuyuki Nishizawa*[‡‡], *Junichi Yamagishi*[‡‡‡,‡‡‡‡], *Minoru Tsuzaki*[†,‡,‡‡‡‡], *Keiichi Tokuda*[†,‡,‡‡‡‡†], *Satoshi Nakamura*[†,‡]

†National Institute of Information and Communications Technology, Japan

‡ATR Spoken Language Communication Research Labs., Japan ‡†Nara Institute of Science and Technology, Japan

‡‡KDDI R&D Labs., Japan ‡‡†Arcadia Inc., Japan ‡‡‡The University of Edinburgh, UK

‡‡‡†Tokyo Institute of Technology, Japan ‡‡‡‡Kyoto City University of Arts, Japan ‡‡‡‡†Nagoya Institute of Technology, Japan

tomoki@is.naist.jp

Abstract

This paper describes the development of a test bed English Text-to-Speech (TTS) system XIMERA at ATR for Blizzard Challenge 2006. The original XIMERA is aimed at constructing very high-quality Japanese TTS. Therefore, several modules are customized for our huge-sized Japanese speech corpora. In order to participate in the Blizzard Challenge 2006, we construct a test bed of English TTS by modifying the original XIMERA without carefully system optimizations. Results of the challenge tell us a current level of our system and clarify points to be improved. This paper also discusses techniques adopted in XIMERA compared with many others.

Index Terms: XIMERA, English TTS, Blizzard Challenge 2006

1. Introduction

The dramatic improvements of Text-to-Speech (TTS) have certainly been caused by the corpus-based approach [1, 2]. That approach has enabled us to construct a TTS system without professional expertise, which is indispensable for constructing the system with consistent and reasonable quality in the rule-based approach [3]. So far, many generic synthesis methods have been established.

In order to better understand different speech synthesis techniques on a common dataset, Blizzard Challenge 2005 was devised in January 2005 [4]. The CMU ARCTIC databases [5] comprised of four speakers were used in the challenge. Each of them consisted of around 1200 phonetically-balanced sentences whose total duration was around one hour. The challenge successfully helped us to better compare several techniques in corpus-based speech synthesis. It is natural to have interests in the same comparison when using larger-sized speech corpus because most of the synthesis techniques are strongly affected by the corpus size. In order to realize it, we recorded a larger-sized US English male speech database at ATR-SLC [6], and provided a 5-hours speech corpus including the CMU ARCTIC subset to Blizzard Challenge 2006.

ATR is one of institutes actively studying corpus-based synthesis techniques such as sample-based synthesis in which a speech waveform is synthesized with acoustic inventories selected from a previously recorded speech corpus. So far, three TTS systems have been developed at ATR, i.e., ν-talk [7], CHATR [8, 9], and XIMERA [10]. The main features of the latest system XIMERA are 1) using a huge-sized speech corpus uttered by a single speaker, e.g., 110-hours corpus of a Japanese male and a 60-hours corpus of a Japanese female, 2) generating target information for segment selection with HMM-based speech synthesis system (HTS) [11, 12], and 3) employing perceptually optimized cost function in segment selection [13, 14]. It has been reported that XIMERA has successfully achieved quite high-quality Japanese synthetic speech [10].

In order to participate in the Blizzard Challenge 2006, we just apply the original XIMERA for Japanese synthesis to English synthesis without carefully system optimizations. Although this system is definitely a test bed, comparing it with other systems is very useful to understand how to improve our system. This paper describes our test bed English XIMERA for the Blizzard Challenge 2006, and then discusses technologies adopted in XIMERA.

The paper is organized as follows. **Section 2** describes the test bed of English XIMERA for the Blizzard Challenge 2006. **Section 3** discusses sample-based synthesis techniques adopted in XIMERA. Finally, we summarize this paper in **Section 4**.

2. Test Bed English XIMERA for Blizzard Challenge 2006

2.1. The Original XIMERA

Figure 1 depicts a block diagram of XIMERA. XIMERA has a generic framework similar to most of TTS systems developed at other institutes. It consists of four major modules such as text processing, target generation, segment selection and waveform generation. More details are described in [10].

XIMERA is aimed at constructing very high-quality Japanese TTS. Therefore, several modules are strongly customized for our huge-sized Japanese speech corpora. Although such a customization causes very high-quality Japanese synthetic voices as shown in [10], it might also cause a lack of flexibility of voice building for various speakers or various languages.

2.2. Development of Test Bed English XIMERA

Considering various points such as dialects, skills, available recording time, and payments, one English male speaker has been selected from 8 male and 9 female candidates. We recorded more than 15-hours English speech data for 18 days over around three months. Details of corpus designing are described in [6].

A test bed of English TTS is constructed as simply as possible by modifying the original XIMERA as follows.

Text processing: A text processor of Festival [15] is just applied to XIMERA without any modifications.

Target generation: Because HTS has already been applied to English synthesis [12], it is straightforward to train HMMs for tar-

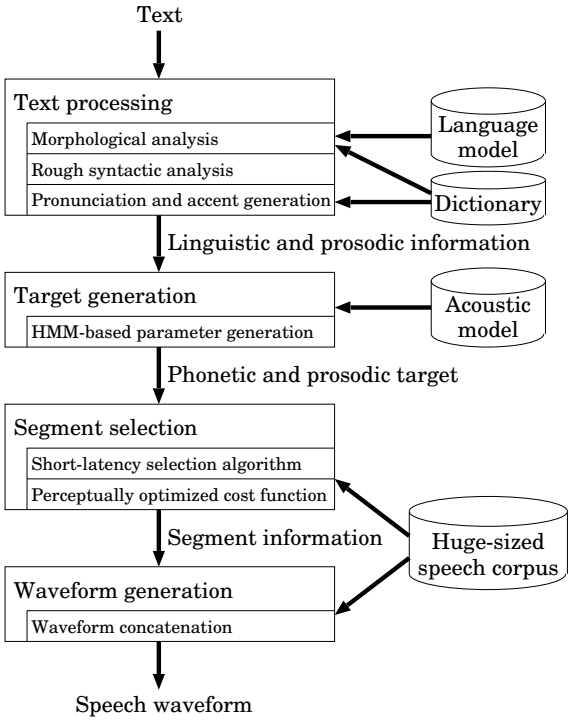

Figure 1: *Block diagram of XIMERA.*

get generation using English speech samples and their transcriptions. Speech parameter generation algorithm considering global variance (GV) [16] is adopted in our test bed system of English XIMERA for improving the quality of generated target parameters. This algorithm generates a speech parameter trajectory that maximizes a product of the likelihood on static and dynamic features and that on the GV. The GV likelihood works as a penalty for the over-smoothing caused by the generalization process, i.e., a reduction of the total variance of the generated parameter trajectory.

Segment selection: In order to simply realize segment selection for English, some sub-cost functions are modified heuristically. For example, we manually define sub-cost tables as shown in **Table 1** that capture the naturalness degradation caused by substitution of phonetic environments. Those settings tend 1) to avoid substitution of phonetic environments of vowel segments often causing formant discontinuities rather than that of consonant segments and 2) to prefer concatenation at boundaries with small power such as C-C to that at boundaries with large power such as V-V often accenting audible discontinuities. These rules are based on our knowledge for Japanese synthesis. Note that the original tables for Japanese XIMERA are much more complex and they were determined by perceptual experiments [17]. Several parameters of cost functions are kept as optimized for our Japanese speaker. A phoneme unit is used as the minimum unit in selection.

Waveform generation: Waveform concatenation works independently of the kind of languages.

2.3. Automatic Voice Building

Voice building is automatically performed using English speech waveforms and *utterance files* automatically generated with Festival provided in the Blizzard Challenge 2006.

F_0 and mel-cepstrum sequences are extracted from speech waveforms with STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum analysis) [18, 19]. And then, in order to perform automatic phoneme segmentation, flat start HMMs for the English speaker are trained as follows:

1. Embedded training of monophone HMMs based on phoneme sequences automatically converted from texts with Festival
2. Copying monophone to triphone HMMs and constructing tied-state triphone HMMs with tree-based clustering
3. Embedded training of the tied-state triphone HMMs
4. Viterbi alignment considering short pause insertion and deletion for refining phoneme sequences
5. Reconstructing tied-state triphone HMMs based on refined phoneme sequences and embedded training of them
6. Viterbi alignment considering short pause insertion and deletion

Three state left-to-right HMMs with a single Gaussian distribution are employed in the above processing. Based on resulting phone segmentation and linguistic and prosodic labels from Festival, HMMs for target generation are trained with HTS [12]. Finally, we constructed a database for segment selection. A main process of that database construction is to train VQ codebooks used for cost calculation.

It takes around one week to finish our voice building process using full database provided in the Blizzard Challenge 2006 whose total duration is around 5 hours if we use single CPU (1.26 GHz Pentium 3). Most of time is spent on context clustering in HTS. If we use multiple CPUs, the voice building time is reduced less than around one day because several processes can be performed simultaneously.

Table 1: *Sub-cost tables on substitution of phonetic environment in test bed English XIMERA. A left table shows sub-costs on phonetic environment substitutions of vowel segments* "V" *and a right one shows those of consonant segments* "C". "Org" *shows preceding/succeeding phoneme categories in the corpus and* "Tar" *shows those in the target phoneme sequence. Note that sub-costs at diagonal elements are set to zeros for successive segments because of no concatenation.*

Curr. V	Tar. V	Tar. C
Org. V	2.0	3.0
Org. C	3.0	1.5

Curr. C	Tar. V	Tar. C
Org. V	1.0	2.0
Org. C	2.0	1.0

2.4. Results of Listening Tests

We submitted two systems as requested by organizers, i.e., *full* which was constructed using 5-hours corpus and *ARCTIC* which was done using the ARCTIC subset. Results show that our system works well as a test bed of English TTS. We plan to improve our system for achieving high-quality English synthetic speech.

3. Discussion of Sample-Based Synthesis Techniques

We discuss techniques adopted in the original XIMERA for clarifying their advantages and disadvantages comparing with other techniques adopted in several TTS systems based on sample-based synthesis [20, 21, 22, 23, 24, 25].

3.1. Speech Database Construction

XIMERA uses huge-sized speech corpora. Currently we have four speech corpora uttered by one Japanese male, one Japanese female, one Chinese female, and one English male. Their corpus sizes are around 110 hours, 60 hours, 20 hours and 15 hours, respectively. Increasing corpus size causes the quality improvements especially in sample-based speech synthesis. However, it also causes some problems, e.g., large voice quality variation [26] and increase of the computational cost for segment selection.

Although manual segmentation is ideal in sample-based speech synthesis, it is a very laborious task. XIMERA employs automatic phone segmentation with HMMs optimized so that an error between resulting segmentation and manually corrected segmentation is minimized in the limited size of speech data [27]. Several techniques such as spectral boundary correction [28] and discriminative HMM training [29] have been proposed for reducing the segmentation error.

3.2. Target generation

XIMERA employs HTS [11, 12] for predicting target parameters for segment selection such as an F_0 contour, phone duration, a power trajectory, and a mel-cepstrum sequence. This method is related to the decision tree-based target prediction [23, 24]. Main advantage of HTS is 1) simultaneous modeling of individual target parameters in the unified framework and 2) generation of smooth target parameter trajectories considering statistics of both static and dynamic features [30]. In contrast to those statistical approaches, selecting sample-based parameters from a speech corpus has also been proposed especially for generating a target F_0 contour [31, 32].

One interesting approach is non-target generation [21]. Segment selection reasonably works directly using linguistic and prosodic labels generated from text processing as the selection targets since target generation is regarded as a process of just converting the kind of features. Compared with such an approach, one of advantages of generating target parameters is controllability of synthetic prosody. Directly controlling each target prosody parameter allows flexible modification of synthetic speech. Another interesting approach is to consider multiple targets [33]. Dealing with prosody variations is one of exciting research themes in TTS.

3.3. Segment selection

There are many attempts at using various units such as not only phone, diphone, and syllable units but also smaller ones, e.g., a half-phone unit [34], an HMM-state level unit [23, 24, 25], and a frame-sized unit [35]. Using shorter units is effective because of an increase of the number of possible unit combinations. However, it makes the computational cost for selection larger. It might also make the possibility of causing audible discontinuities larger due to the difficulty of accurately detecting them with existing acoustic distance measures [36, 37]. Considering those advantages and disadvantages, XIMERA uses a half phoneme as the minimum unit for Japanese synthesis under knowledge-based concatenation rules such as avoiding concatenation at several phoneme combinations often causing audible discontinuities.

There are several approaches for reducing the computational cost such as caching concatenation costs [38] and pre-selection [39]. Tree-based unit clustering adopted in several systems [22, 23, 24, 25] is also an effective way. XIMERA employs preselection based on a target cost and VQ-based cost calculation for further reducing the computational cost. Moreover, a short latency unit selection algorithm [40] is used for realizing a fast response of synthetic speech. This on-the-fly selection algorithm allows presenting a synthetic speech while searching an appropriate segment sequence. Consequently, a response time is less than around 800 ms even if using our huge-sized speech corpus.

It is essential to use a cost sensitively capturing the naturalness degradation in segment selection. XIMERA uses cost functions optimized based on perceptual experiments. Acoustic parameters and linguistic features are converted to individual sub-costs capturing the naturalness degradation caused by individual factors based on sub-cost functions [13]. And then, those sub-costs are integrated to a cost capturing the naturalness degradation of a selected segment sequence based on an integrated cost function [14]. Perceptual optimization is a solid way for improving the naturalness of synthetic speech. It also enables us to estimate a perceptual score such as mean opinion score (MOS) from a cost [14, 41]. However, it is quite laborious and time-expensive task. Moreover, because the number of stimuli to be evaluated is limited, it is necessary to simplify the cost functions. Recently, it has been reported that statistically defined cost functions slightly outperforms perceptually optimized ones [42]. It seems promising to perceptually optimize only simple factors such as weights for target and concatenation costs after statistically defining cost functions.

3.4. Waveform generation

Since the desired waveform segments with target prosodic parameters are not always in the corpus, other segments whose prosodic parameters close to the target ones need to be used instead. Directly concatenating such waveform segments often causes the naturalness degradation due to prosody distortion. Sophisticated signal processing techniques such as TD-PSOLA (Time-Domain Pitch-Synchronous OverLap-Add) [43], Harmonic plus Noise Model (HNM) [44], and STRAIGHT [19] are useful for alleviating it. However, even such state-of-the-art techniques might cause other artificial sounds due to essential problems of speech analysis: difficulties of estimating an accurate vocal tract response from sparse frequency components observed at only F_0 harmonic points. Beutnagel et al. [45] reported waveform concatenation outperforms HNM-based prosody modification.

The effectiveness of prosody modification seems to be strongly affected by the corpus size. Our experimental evaluations using STRAIGHT-based prosody modification show the following results. The prosody degradation is a dominant factor of the naturalness degradation when the corpus size is small because it is difficult to find waveform segments acceptably realizing the target prosody. An increase of the corpus size effectively decreases prosody degradation. Although it also alleviates artificial sounds caused by signal processing due to a decrease of the modification rate, the quality improvements of prosody modification are less than those of waveform concatenation. Consequently, waveform concatenation outperforms prosody modification when using large-sized speech corpus. XIMERA usually employs waveform concatenation since the corpus size is enough large.

4. Conclusions

This paper described a test bed English Text-to-Speech (TTS) system XIMERA from ATR for Blizzard Challenge 2006. It was developed by just applying the original XIMERA for Japanese synthesis to English synthesis without any optimizations. We will

improve our system based on useful results of the challenge for achieving high-quality English synthetic speech.

5. References

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. *Proc. of ICASSP*, pp. 679–682, 1988.

[2] T. Hirokawa. Speech synthesis using a waveform dictionary. *Proc. EUROSPEECH*, pp. 140–143, 1989.

[3] D.H. Klatt. Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, vol. 82, no. 3, pp. 737–793, 1987.

[4] A.W. Black and K. Tokuda. The Blizzard Challenge 2005: evaluating corpus-based speech synthesis on common datasets. *Proc. of Interspeech*, pp. 77–80, 2005.

[5] J. Kominek and A. Black. The CMU ARCTIC speech databases for speech synthesis research. *Technical Report*, CMU-LTI-03-177, Carnegie Mellon University, 2003.

[6] J. Ni, T. Hirai, and H. Kawai. Constructing a phonetic-rich speech corpus while controlling time-dependent voice quality variability for English speech synthesis. *Proc. ICASSP*, pp. 881–884, 2006.

[7] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura. ATR ν-talk speech synthesis system. *Proc. ICSLP*, pp. 483–486, 1992.

[8] A.W. Black and P. Taylor. CHATR: a generic speech synthesis system. *Proc. COLING*, pp. 983–986, 1994.

[9] W.N. Campbell. CHATR: a high-definition speech re-sequencing system. *Proc. Joint Meeting of ASA and ASJ*, pp. 1223–1228, 1996.

[10] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda. XIMERA: a new TTS from ATR based on corpus-based technologies. *Proc. of 5th ISCA Speech Synthesis Workshop (SSW5)*, pp. 179–184, 2004.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Proc. of EUROSPEECH*, pp. 2347–2350, 1999.

[12] K. Tokuda, H. Zen, and A.W. Black. An HMM-based speech synthesis system applied to English. *Proc. of IEEE 2002 Workshop on Speech Synthesis*, 2002.

[13] T. Toda, H. Kawai, and M. Tsuzaki. Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis. *Proc. ICASSP*, pp. 657–660, 2004.

[14] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano. An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis. *Speech Communication*, vol. 48, no. 1, pp. 45–56, Jan. 2006.

[15] *The Festival speech synthesis system*. http://www.cstr.ed.ac.uk/projects/festival/

[16] T. Toda and K. Tokuda. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *Proc. of INTERSPEECH*, pp. 2801–2804, Sep. 2005.

[17] H. Kawai and M. Tsuzaki. Acoustic measures vs. phonetic features as predictors of audible discontinuity in concatenative speech synthesis. *Proc. ICSLP*, pp. 2621–2624, 2000.

[18] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity. *Proc. EUROSPEECH*, pp. 2781–2784, 1999.

[19] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: possible role of a repetitive structure in sounds. *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[20] A.K. Syrdal, C.W. Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K-S. Lee, and M.J. Makashay. Corpus-based techniques in the AT&T NextGen synthesis system. *Proc. of ICSLP*, vol. 3, pp. 410–415, 2000.

[21] M. Chu, H. Peng, H. Yang, and E. Chang. Selecting non-uniform units from a very large corpus for concatenative speech synthesizer. *Proc. ICASSP*, pp. 785–788, 2001.

[22] A.W. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. *Proc. EUROSPEECH*, pp. 601–604, 1997.

[23] R.E. Donovan and P.C. Woodland. A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, vol. 13, no. 3, pp. 223–241, 1999.

[24] E. Eide, A. Aaron, R. Bakis, P. Cohen, R. Donovan, W. Hamza, T. Mathes, M. Picheny, M. Polkosky, M. Smith, and M. Viswanathan. *Recent improvements to the IBM trainable speech synthesis system*. *Proc. ICASSP*, pp. 708–711, 2003.

[25] X. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, J. Liu and M. Plumpe. Whistler: a trainable text-to-speech system. *Proc. of ICSLP*, pp. 2387–2390, 1996.

[26] H. Kawai and M. Tsuzaki. A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis. *Proc. IEEE 2002 Workshop on Speech Synthesis*, 2002.

[27] H. Kawai and T. Toda. An evaluation of automatic phone segmentation for concatenative speech synthesis. *Proc. ICASSP*, pp. 677–680, 2004.

[28] Y-J. Kim and A. Conkie. Automatic segmentation combining an HMM-based approach and spectral boundary correction. *Proc. ICSLP*. pp. 145–148, 2002.

[29] Y. Wu, H. Kawai, J. Ni, and R. H. Wang. Discriminative training and explicit duration modeling for HMM-based automatic segmentation. *Speech Communication*, vol.47, no. 4, pp. 397–410, Dec. 2005.

[30] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. of ICASSP*, pp. 1315–1318, 2000.

[31] M. Isogai and H. Mizuno. A new F_0 contour control method based on vector representation of F_0 contour. *Proc. EUROSPEECH*, pp. 727–730, 1999.

[32] A. Raux and A.W. Black. A unit selection approach to F_0 modeling and its application to emphasis. *Proc. of ASRU*, pp. 700–705, 2003.

[33] I. Bulyko and M. Ostendorf. Efficient integrated response generation from multiple targets using weighted finite state transducers. *Computer Speech and Language*, vol. 16, No. 3–4, pp. 533–550, 2002.

[34] A. Conkie. Robust unit selection system for speech synthesis. *Joint Meeting of ASA, EAA, and DAGA*, 1999. http://www.research.att.com/ ttsweb/tts/pubs.php

[35] T. Hirai and S. Tenpaku. Using 5 ms segments in concatenative speech synthesis. *Proc. of 5th ISCA Speech Synthesis Workshop (SSW5)*, pp. 37–42, 2004.

[36] E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Trans. Speech and Audio Processing*, vol. 9, No. 1, pp. 39–51, 2001.

[37] Y. Stylianou and A.K. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. *Proc. ICASSP*, pp. 837–840, 2001.

[38] M. Beutnagel, M. Mohri, and M. Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. *Proc. EUROSPEECH*, pp. 607–610, 1999.

[39] A. Conkie, M. Beutnagel, A.K. Syrdal, and P.E. Brown. Preselection of candidate units in a unit selection-based text-to-speech synthesis system. *Proc. ICSLP*, vol. 3, pp. 279–282, 2000.

[40] N. Nishizawa and H. Kawai. A short-latency unit selection method with redundant search for concatenative speech synthesis. *Proc. ICASSP*, pp. 757–760, 2006.

[41] M. Chu and H. Peng. An objective measure for estimating MOS of synthesized speech. *Proc. EUROSPEECH*, pp. 2087–2090, 2001.

[42] Y. Zhao, P. Liu, Y. Li, Y. Chen, and M. Chu. Measuring target cost in unit selection with KL-divergence between context-dependent HMMs. *Proc. ICASSP*, pp. 725–728, 2006.

[43] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, Vol. 9, No. 5–6, pp. 453–467, 1990.

[44] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[45] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T Next-Gen TTS system. *Joint Meeting of ASA, EAA, and DAGA*, 1999. http://www.research.att.com/ ttsweb/tts/pubs.php