THE IMPACT OF SPEECH RECOGNITION ON SPEECH SYNTHESIS

Mari Ostendorf and Ivan Bulyko

Department of Electrical Engineering
University of Washington, Seattle, WA 98195.
{mo,bulyko}@ssli.ee.washington.edu

ABSTRACT

Speech synthesis has changed dramatically in the past few years to have a corpus-based focus, borrowing heavily from advances in automatic speech recognition. In this paper, we survey technology in speech recognition systems and how it translates (or doesn't translate) to speech synthesis systems. We further speculate on future areas where ASR may impact synthesis and vice versa.

1. INTRODUCTION

In the last decade, there has been a paradigm shift in speech synthesis work such that both commercial and research systems are predominantly based on data-driven or corpus-based techniques. The emphasis is on engineering techniques – including optimization of cost functions, statistical models and signal processing – more than on linguistic rule development. Many of the current text-to-speech (TTS) techniques, as well as the philosophy, are borrowed from automatic speech recognition (ASR). There is also a growing interest in limited-domain synthesis, driven at least in part by the success of limited-domain speech understanding systems as well as the need for such systems to have high quality speech generation for telephony applications.

There is no question that these changes have greatly advanced the state of the art in speech synthesis, particularly in the sense of much more natural sounding voices. Furthermore, there are many promising avenues for advances from leveraging other aspects of ASR technology. However, the shift has had a cost in the reduced amount of scientifically oriented research, both that aimed at the end goal of text-to-speech synthesis but also studies of speech communication that use synthesis as a tool, since non-parametric synthesizers do not provide the controls needed for such studies. In addition, there has been surprising little transfer of speech synthesis technology back to the recognition community.

The goal of this paper is to survey recent work associated with this data-oriented trend, showing the connections to speech recognition as well as missing links (i.e. where ASR technology can be taken further for synthesis applications) and potential pitfalls (i.e. where it fails). In Section 2, we review the key elements of speech recognition technology, and in Section 3 we survey ways in which this technology is used in speech synthesis. Section 4 outlines limitations of the "technology transfer," including examples of where synthesis and ASR techniques differ in important ways. Finally, in Section 5, we speculate on new directions for using recognition technology in synthesis, as well as possibilities for transferring technology back.

2. SPEECH RECOGNITION

Speech recognition itself underwent a paradigm shift two decades ago, when statistical models began to be broadly adopted. Three key properties were important in their success. First, statistical models provided a means of "ignorance modeling," i.e. representing broad distributions (or rough characterizations) of phenomena for which linguistic rules were not yet known or speech knowledge was incomplete. Second, data-driven learning techniques could be applied to structure and rule learning, not simply parameter learning, and in that sense they provide a much less costly alternative to human labor for writing rules. Finally, the idea of "delayed decisions" was vital for improving performance, as systems moved from using intermediate hard decisions about phonemes or phonetic features to an integrated word recognition search that entertained multiple word hypotheses at any point in time. (It is interesting to note that, while the ASR justification for delayed decisions is grounded in optimization theory, similar ideas also appear in psycholinguistic models of human spoken language processing [76].) While "ignorance modeling" has little appeal in the synthesis community, the other two properties are clearly recognized as valuable.

Since that time, ASR technology has advanced considerably, from success only on small vocabulary, speaker-dependent, read-speech tasks to high performance on large vocabulary, speaker-independent naturally occurring speech. ASR is still far from a solved problem, with word error rates of 13%, 24% and 36% on broadcast news, conversational speech and meeting speech, respectively, and higher error rates under noisy conditions [44]. However, performance on broadcast news transcription is good

enough for information retrieval applications, and many limited-domain commercial ASR systems are deployed in telephony and embedded system applications. This success is widely attributed to the data-driven approach, as well as to the international competitions that have become traditional in the ASR community over the past decade. Of course, Moore's law has also been an important facilitator: increasing computing power made it possible to train on larger corpora and search over larger word hypothesis spaces using models with increased complexity.

The basic components in most large vocabulary speech recognition systems include: mel-cepstral speech analysis, hidden Markov modeling of acoustic subword units (typically context-dependent phones), clustering techniques for learning parameter tying, n-gram models of word sequences, a beam search to choose between candidate hypotheses, adaptation of acoustic models to better match a target speaker, and multi-pass search techniques to incorporate adaptation as well as models of increased complexity. A detailed review of these components is beyond the scope of this paper, but we will briefly describe key elements that are built on in speech synthesis. For more details, readers are referred to [79, 36, 33]. We also note that many of the same techniques are also used in speaker recognition [50].

Mel-cepstral feature extraction is used in some form or another in virtually every state-of-the-art speech recognition system. Using a rate of roughly 100 frames/second, speech windows of 20-30ms are processed to extract cepstral features, then normalized to compensate for channel mismatch and speaker variability. While there are several ways of computing mel-cepstra, a popular approach is to use a mel-spaced bank of filters to obtain a vector of log energies, to which is applied a cosine transform [32]. The cepstral features are typically augmented with their first- and second-order derivatives to form observation vectors. Signal processing may also include subsequent linear transforms to reduce the vector dimension.

The *hidden Markov model (HMM)* is also widely used in speech recognition systems. The HMM represents acoustic variability in two ways: a hidden, discrete Markov state sequence characterizes temporal variability; and a state-dependent observation distribution characterizes spectral variability, typically using Gaussian mixtures. Observations are assumed to be conditionally independent given the state sequence; hence, HMMs are sometimes described as a piecewise-constant generative model (i.e. the mean is piecewise-constant, given the state sequence) and thus a poor model of speech dynamics. However, the use of derivatives in the observation space (though violating the conditional independence assumption) effectively counteracts this problem, both in recognition and, as we shall see, in synthesis. Although there have been many alternative statistical approaches to acoustic modeling investigated since the development of HMMs (e.g. [46, 41, 9]), for the most part, these can be thought of as extensions of HMMs.

An important tool for building complex ASR systems is *distribution clustering* for parameter tying. A popular approach is maximum likelihood decision-tree-based clustering, which is a divisive clustering strategy that uses linguistically motivated candidate splitting questions to provide a mechanism for generalizing to unseen data. Different methods are used depending on the parameter tying assumptions, e.g. at the distribution level [80] vs. at the mixture component level [22]. Clustering can also be used to learn HMM topologies [47] and pronunciation models [27, 55]. Clustering is used extensively in other parts of an ASR system as well, from adaptation to language modeling.

Another component in virtually every large vocabulary recognition system is the *n-gram model* for representing word sequence probabilities. In general, an n-gram model describes a symbol sequence as an $(n-1)$-th order Markov process, and can be used for phones (as in language identification), words (as in ASR), or prosodic events (as in TTS). A key issue in n-gram modeling is estimation of low frequency events using smoothing and/or back-off techniques; see [17] for a good survey. While much research has sought to improve over this simple language model, only small gains have been achieved so far [51, 30].

Given the HMM and n-gram statistical models of acoustic and language components of speech, respectively, speech recognition can be formulated as a *search* problem: find the maximum probability word sequence given the acoustic observations. This means computing the likelihoods of and comparing all possible word sequences, which would be impractical except for the assumptions of conditional independence of observations and word history given local context. These assumptions make it possible to keep track of the best path and score only at the most recent previous word-state. The efficient implementation of the optimal search involves dynamic programming and is typically referred to as the Viterbi algorithm, which is borrowed from communication theory. As vocabulary sizes and model complexity has continued to grow, the full search can still be impractical and various techniques are used to improve efficiency, including pruning (beam search) and tree structured lexicons [43] and weighted finite-state transducers [40].

An important component of high-performance, large vocabulary ASR systems is *speaker adaptation*, i.e. changing the parameters of a speaker-independent acoustic model to better match the speech from a target speaker. Early work and some commercial systems use supervised adaptation, or speaker enrollment, where the word transcriptions for the target speaker are known. The current research emphasis is on unsupervised adaptation, including both on-line and batch methods. There is a large body of work in this area, involving a variety of techniques, but the most influential is

an approach involving a linear transformation (rotation) of the means and optionally variances of the observation distributions. The transformations are trained to maximize the likelihood of data from the new speaker, hence the name maximum likelihood linear regression (MLLR) adaptation [39]. Multiple transformations are often used and are associated with different sub-word models via clustering. Many variations of MLLR have since been developed; of most relevance to synthesis is probably the mixture transformation model [21].

Another important element of modern speech recognition systems is the significant attention to system engineering. While the engineering aspects are often downplayed as "uninteresting", they have a major impact on the performance of ASR systems and one might argue that the careful experimentation behind good system engineering is also a characteristic of good science.

3. ASR IN SYNTHESIS

Virtually every one of the above techniques is now incorporated in one or more concatenative speech synthesis systems. The more philosophical elements, including participation in competitive evaluations, are also playing a role in the synthesis community. Below we survey areas in which synthesis is benefiting from ASR technology, focusing on a few significant areas rather than attempting to provide an exhaustive review.

3.1. Speech Synthesis as a Search Problem

Probably the first idea from ASR that had a major impact on synthesis was the dynamic programming search algorithm, used for selecting units to concatenate from a large corpus. Pioneering work in this area was at ATR [54]. The basic idea is that, while it is sufficient to have one instance of each unit in the database (e.g. all possible phone transitions when using diphones) to be able to synthesize any possible input, the output quality of a concatenative synthesizer can be improved if there are multiple instances of each unit to choose from, especially when automatic methods are used to segment and annotate units. Having a larger number of units to choose from during synthesis makes it possible to find units that are more suitable for a given phonetic and prosodic context, hence reducing the amount of additional signal processing.

Selection of variable size units from large single-speaker speech databases usually involves minimizing distortion introduced when selected units are modified and concatenated, e.g. [54, 35, 34, 19, 24, 7]. This distortion is represented in terms of two cost functions: 1) a *target* cost, which is an estimate of the difference between the database unit u_i and the target t_i, and 2) a *concatenation* cost, which is an estimate of the quality of concatenation of two units u_{i-1} and u_i. The task of unit selection, i.e. finding a sequence of database units that minimize the sum of the target and concatenation costs, is implemented with a Viterbi search. The unit selection search is, in some sense, a reverse of the decoding process in speech recognition. Instead of finding a word sequence that best explains acoustic input as in ASR, synthesis is about finding a sequence of acoustic units to optimally represent a given word sequence. However, there are nice parallels that make the algorithms transfer, i.e. the concatenation cost is like a state transition score (though more powerful) and the target cost is like a state-dependent observation probability. In both cases, context-dependent phonetic subword units are concatenated to form words, though in synthesis the "context" is likely to include lexical stress and possibly prosodic variables.

As described above, the steps of recognizing phones from acoustics and choosing words to match the phone sequence are coupled in ASR in a joint search routine that effectively delays assignment of phones to the acoustic stream until it is clear that these phones can form a legitimate word sequence. This idea of a delayed decision is used in current unit selection algorithms for TTS, but it can be expanded to include other stages of the synthesis problem as well. For example, instead of predicting target prosody first and then searching for units to match that target, alternative prosodic realizations of a sentence can be evaluated jointly with the unit sequence [13], so prosody prediction is effectively a "soft" decision. This approach takes advantage of the fact that one can convey essentially the same meaning with prosodically different yet perceptually acceptable realizations of the same utterance, as evidenced by the variability observed in different readings of the same text [52]. Taking this idea one step further, one can allow variation in the word sequence and jointly optimize the wording, prosody and unit sequence [15]. Of course, with every level of processing where alternatives are allowed, the search space increases. Hence, the "synthesis as search" view is probably most useful for limited-domain applications.

3.2. ASR Tools for Annotation

In corpus annotation, speech recognition offers automatic phonetic alignment and decision tree based clustering for grouping speech segments according to their context. Having the ability to reduce (or avoid) manual preparation of speech corpora makes it possible to take advantage of large databases and reduces the cost of developing new voices.

While the general methodology for automatic segmentation (forced Viterbi alignment) is common in both synthesis and recognition, there are differences in specific approaches. The output quality of a concatenative speech synthesizer relies heavily on the accuracy of segment boundaries in the speech database, however, speech recognition systems are typically not optimized for the phonetic

segmentation task. To reduce forced alignment errors researchers have investigated automatic error correction methods [48], edge detector outputs as features [73], and non-uniform HMM topologies (1-state HMM for fricatives and 2-state HMM for stops) [4].

Decision tree clustering has also been used in speech synthesis to define the inventory of units (similar to speech recognition) either by tying HMM parameters to maximize likelihood of the training data [24, 4] or by minimizing the average distance between units in each cluster [10]. While ASR includes only spectral envelope features in the objective function, synthesis may include acoustic/prosodic information such as pitch and duration. A more important difference, however, is that ASR uses the combined statistics from all the data in a cluster (leaf node of the tree), whereas TTS uses specific speech instances in the clusters. Further, "outliers" and overly "common" units are often removed from their clusters [10]. Due to the binary nature of decision tree splitting and the use of linguistic features to choose between units, the resulting clusters may overlap in the acoustic space covered even though they do not overlap in terms of set membership. In synthesis, such overlapping clusters may be desirable for finding the best unit sequence, to the point of introducing sharing of some units in multiple clusters [12].

3.3. Speech Models

Hidden Markov models have been used more directly in speech synthesis in two ways: as a model on which to assess or reduce target and concatenation costs, and as a generative model for the actual synthesis process.

In concatenative speech synthesis, the output quality relies on the system's ability to avoid or reduce the artifacts introduced when speech segments are spliced together. This motivates the use of models of spectral dynamics, especially for short-term dynamics at the unit boundaries. To this end, HMMs have been applied in synthesis for a) assessing the smoothness of joins between pairs of units formulated by concatenation costs [25, 16], and b) smoothing spectral discontinuities at unit boundaries in the output waveform [49].

Initial work on using HMMs as a generative model was dismissed, and in fact used anecdotally to demonstrate that HMMs were not good models of speech by using them to generate speech (a sequence of spectral parameters) either through random sampling or with state distribution means. The quality of both HMM recognition and synthesis, however, relies heavily on the ability of an HMM to capture spectral dynamics. In [69, 71], an algorithm is described for generating speech parameters from continuous mixture HMMs with dynamic features. It was demonstrated that synthetic speech obtained without dynamic features exhibits perceptible discontinuities, while with dynamic features the output speech is much smoother. This work has been extended to include source characteristics (fundamental frequency) [70], and other statistical models have also been investigated [26, 16].

3.4. Statistical models for prosody prediction

Work in using corpus-based machine learning techniques for speech synthesis was pioneered by Hirschberg and colleagues [74, 31], using decision trees for symbolic prosody prediction (prosodic prominence and phrasing). Since then, a wide variety of data-driven techniques have been explored for predicting symbolic prosodic events, including (for example) combinations of decision trees and Markov process models [52], hidden Markov models [11], and transformational rule-based learning [28]. Prosody prediction algorithms for TTS rely on text normalization, tagging and sometimes parsing to provide features as input to the models. While data-driven techniques have long dominated work in tagging and parsing, text normalization (expansion of abbreviations, numbers, etc.) was mostly based on ad hoc rules. Recently, however, the use of trainable techniques (including n-gram models and decision trees) has also been extended to text normalization [61].

Corpus-based techniques have also played an important role in predicting continuous prosodic parameters, particularly duration and intonation. While regression trees (a variant of decision trees for predicting continuous variables) are used for both applications, more sophisticated models have emerged. For example, duration modeling using the sums-of-products model [72] has been extended to incorporate data transformations in the data-driven estimation process [6]. Corpus-based models of intonation include those that are explicitly tied to a linguistically motivated symbolic representation (e.g. ToBI tones) using a statistical dynamical system model (an HMM extension) [53] or iterative analysis-resynthesis training [60], as well as more explicitly data-driven approaches [68, 42].

4. LIMITATIONS

Fundamentally recognition and synthesis are different problems. ASR must account for speaker variability and ignore speaker-specific details that are important for synthesis quality. For synthesis, it is sufficient to generate a single rendition of a sentence with high quality; even some sources of intra-speaker variability can be ignored by the model, particularly that associated with spontaneous speech (e.g. disfluencies). Since recognition involves discriminating between different classes, discriminative learning techniques are relevant and are becoming more pervasive in speech recognition research [77, 18]. Synthesis, on the other hand, requires a descriptive model – one that fully characterizes the speech signal – in which case the maximum likelihood (ML) training criterion is most appropriate.

Indeed, research applying both types of criteria for learning structural dependencies between observations in a buried Markov model (HMM extension) show that an ML objective hurts recognition [9] but is preferred for synthesis [16].

One area where speech synthesis and recognition clearly depart is in signal processing. ASR mostly ignores prosody – which is vital to synthesis – and glottal source characteristics more generally. As described above, recognition tasks rely on mel-cepstral processing, even in speaker recognition where the source characteristics are likely to be relevant. In speech synthesis, harmonic models of speech are popular, e.g. [63, 45], because they tend to give high quality speech in conjunction with duration and fundamental frequency modifications needed for prosody control. Further, mel-cepstral distances are shown in one study to be among the worst predictors of audible discontinuities in perceptual experiments [38], though other work shows that the performance difference relative to the best case (a Kullback-Liebler spectral distance) is not that large [64].

Another area where speech recognition technology is limited, again because of the poor representation of source characteristics, is in voice conversion (or transformation). Voice conversion is an important technology for delivering a variety of voices using corpus-based synthesizers, whether concatenative or HMM-based. ASR adaptation techniques have long been used for speech synthesis voice conversion [1], and for spectral conversion (i.e. ignoring source characteristics) it makes good sense as the two problems are essentially the same. Given a large amount of data for training the initial model (typically speaker-independent for ASR, speaker-dependent for TTS), and a small amount of data from a target speaker, find a mapping to transform the source model to match the target speaker. Not surprisingly, MLLR has been applied to speech synthesis [66], and one can show that the Gaussian mixture mapping used in synthesis [65, 37] is equivalent to the mixture MLLR approach [21]. However, these techniques provide only a partial solution, since they fail to capture the prosodic idiosyncrasies and glottal source characteristics of the speaker. There is some work showing that one can apply the same adaptation techniques to both spectral and source parameters [3, 67], but these operate at essentially the frame level and much more research is needed to effectively adapt the suprasegmental characteristics of a speaker.

5. SPECULATIONS

There is some concern that there has been an over-emphasis on ever larger data sets and faster computing (consistent with ASR) at the expense of fundamental research. The same concern is raised for ASR, of course, leading some to say that ASR has reached a local optimum that will require significant change to break out of. Whether or not this is the case for ASR, corpus-based synthesis is clearly not at a "plateau" yet, being still in its early stages, and there are significant gains yet to be had by pursuing the current research trends. Furthermore, future advances in ASR may provide new avenues for use of ASR techniques. However, it would be a mistake to completely abandon the more fundamental speech modeling questions traditionally the domain of synthesis research. We argue here that parametric and data-driven techniques can and should co-exist, and furthermore that advances in data-driven parametric synthesis will eventually impact speech recognition.

5.1. ASR and Parametric Synthesis

Although concatenative synthesis systems are thought by many to be higher quality than parametric synthesis systems, there are many demonstrations of copy synthesis that show the potential for very high quality from the parametric approach. The parametric approach tends to be better suited to scenarios with a large variety of voices; generation of speech with singing, whispering or laughing; and small footprint or low power applications. A key limitation of parametric synthesizers has been the high cost of rule development, but ASR tools offer the potential to change this, in combination with advances in estimating articulatory and/or formant and glottal source control parameters from speech.

Traditionally, formant synthesizers have been implemented with manually derived rules that specify the formant trajectory, but it is possible to create data-driven formant synthesizers. For example, HMMs have been used for formant synthesis [2], similarly to [69, 70] but using formant frequencies and bandwidths as features. To make this possible it is crucial that the formant extraction is done as part of the HMM parameter training and not done independently.

5.2. Can Synthesis Impact Recognition?

In our own work, synthesis already has had an influence on ASR, including the use of text normalization tools [61] for language modeling [56] and the use of word/syllable structure in context conditioning for HMM distribution clustering [24, 57]. Other ASR researchers are looking to use formants as hidden features in the search space in the so-called hidden dynamical systems [20]. We also hypothesize that acoustic modeling improvements that lead to better phonetic segmentation for TTS applications will generalize to infrequent contexts better and hence offer improvements to ASR. However, none of these examples are broadly adopted in ASR. While there will probably be more such examples, we see two main areas where synthesis may have a more significant impact on recognition in the future: in the use of prosody for speech transcription and understanding, and in speaker recognition.

There is growing evidence that prosody may be useful for speech recognition, including recent work by Stolcke incorporating a pause language model [62] and in context-dependent acoustic modeling [58]. However, prosody is clearly important for applications where more than just a word transcription is needed. One important area is in spoken dialog systems, where prosody can be useful for dialog act recognition as well as in parsing [5]. A second area is annotation of structure in speech, including punctuation, disfluencies and topic segmentation [59]. While still in early stages, representation of emotion in speech is also of interest for augmented transcriptions.

Though prosody played a role in early speaker recognition systems [29], prosodic features have mostly been abandoned for cepstral features in the context of Gaussian mixture models. Recently, however, there is renewed interest in prosody in the area of text-independent speaker recognition with extended speech sources, where modeling of high level patterns of word usage and speaking style are possible. In this context, it was shown that prosodic features can provide added information on top of the standard cepstral cues in [75]. Further gains may be possible by investigating synthesis-style signal processing that better captures source characteristics.

Much of the work on prosody modeling for recognition applications has not built on synthesis models, in part because of the inherent difference in speaker-independent recognition vs. speaker-specific generation. However, as the prosody models become more speaker-specific (for speaker identification) or speaker-adaptive (for recognition), we anticipate that prosody modeling will be a fruitful area for collaboration between synthesis and recognition researchers.

Acknowledgments

This authors wish to thank Alex Acero for many useful discussions as well as input on an earlier version of this paper. This work was supported by DARPA grant no. N660019928924 and by the National Science Foundation under Grant No. IIS-9528990. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

6. REFERENCES

[1] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, "Voice conversion through vector quantization," *Proc. ICASSP*, pp. 1:655-658, 1988.

[2] A. Acero, "Formant analysis and synthesis using hidden Markov models," *Proc. Eurospeech*, 1:1047-1050, 1999.

[3] L. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum," *Proc. ICASSP*, 3:1347-1350, 1997.

[4] H. Hon, A. Acero, X. Huang, J. Liu and M. Plumpe, "Automatic generation of synthesis units for trainable text-to-speech systems," *Proc. ICASSP*, 1:293-296, 1998.

[5] A. Batliner *et al.*, "Whence and whither prosody in automatic speech understanding: a case study," *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 3-12, 2001.

[6] J. Bellegarda, K. Silverman, K. Lenzo and V. Anderson, "Statistical Prosodic Modeling: From Corpus Design to Parameter Estimation," *IEEE Trans. Speech and Audio Processing*, **9**(1):52-66, 2001.

[7] 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*, 1:18-24, 1998.

[8] M. Beutnagel, M. Mohri and M. Riley, "Rapid unit selection from a large speech corpus for concatenative speech synthesis," *Proc. Eurospeech*, 2:607-610, 1999.

[9] J. Bilmes, "Buried Markov Models for speech recognition," *Proc. ICASSP*, 2:713-716, 1999.

[10] A. Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis," *Proc. Eurospeech*, 2:601-604, 1997.

[11] A. Black and P. Taylor, "Assigning phrase breaks from part-of-speech sequences," *Computer, Speech and Language*, **12**(2):99-117, 1998.

[12] I. Bulyko, *Flexible speech synthesis using weighted finite-state transducers*, University of Washington, Ph.D. Dissertation, Electrical Engineering, 2002.

[13] I. Bulyko and M. Ostendorf, "Joint Prosody Prediction and Unit Selection for Concatenative Speech Synthesis," *Proc. ICASSP*, 2:781-784, 2001.

[14] I. Bulyko and M. Ostendorf, "Unit selection for speech synthesis using splicing costs with weighted finite state transducers," *Proc. Eurospeech*, 2:987-990, 2001.

[15] I. Bulyko and M. Ostendorf, "Efficient Integrated Response Generation from Multiple Targets using Weighted Finite State Transducers,," *Computer Speech and Language*, to appear, July 2002.

[16] I. Bulyko, M. Ostendorf, and J. Bilmes, "Robust splicing costs and efficient search with BMM models for concatenative speech synthesis," *Proc. ICASSP*, 1:461-464, 2002.

[17] S. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, **13**(4):359-394, 1999.

[18] W. Chou, "Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition," *Proc. IEEE*, **88**(8):1201-1223, 2000.

[19] A. Cronk and M. Macon, "Optimized Stopping Criteria for Tree-Based Unit Selection in Concatenative Synthesis," *Proc. ICSLP*, 1:680-683, 1998.

[20] L. Deng, "A Dynamic, Feature-Based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition," *Speech Communication*, 24:299-323, 1998.

[21] V. Diakoloukas and V. Digalakis, "Maximum likelihood stochastic-transformation adaptation of hidden Markov models," *IEEE Trans. Sp. and Audio Proc.,* **7**(2):177-187, 1999.

[22] V. Digalakis, P. Monaco and H. Murveit, "Genones: Generalized Mixture Tying in Continuous Hidden Markov Model Based Speech Recognizers," *IEEE Tran. on Speech and Audio Processing*, **4**(4):281-289, 1996.

[23] R. Donovan, "Segment preselection in decision-tree based speech synthesis systems," *Proc. ICASSP*, 2:937-940, 2000.

[24] R. Donovan and P. Woodland, "A hidden Markov-model-based trainable speech synthesizer," *Computer Speech and Language*, **13**(3):223-241, 1999.

[25] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers," *Proc. ESCA Workshop on Speech Synthesis*, 2001.

[26] M. Eichner *et al.,* "Speech synthesis using stochastic Markov graphs," *Proc. ICASSP,* 2:829-832, 2001.

[27] E. Eide, "Automatic modeling of pronunciation variation," *Proc. Eurospeech*, 1:451-454, 1999.

[28] C. Fordyce and M. Ostendorf, "Prosody prediction for speech synthesis using transformational rule-based learning," *Proc. ICSLP,* 3:843-846, 1998.

[29] S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," *Speech Commun.*, 5:183-197, 1986.

[30] J. Goodman, "A bit of progress in language modeling," *Computer Speech and Language,* **15**(4):403-434, 2001.

[31] J. Hirschberg, "Pitch accent in context: predicting intonational prominence from text," *Artificial Intelligence*, 63:305-340, 1993.

[32] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland, *The HTK Book (Version 2.0)*. ECRL, 1995.

[33] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, NJ, 2001.

[34] X. Huang *et al.,* "Whistler: a trainable text-to-speech system," *Proc. ICSLP,* 4:169-172, 1996.

[35] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *Proc. ICASSP*, 1:373-376, 1996.

[36] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.

[37] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP,* 1:285-288, 1998.

[38] E. Klabbers and R. Veldhuis, "Reducing Audible Spectral Discontinuities," *IEEE Trans. Speech and Audio Processing*, **9**(1):39-51, 2001.

[39] C. J. Leggetter and P. Woodland, "Speaker adaptation using maximum likelihood linear regression," *Computer Speech and Language*, **9**(2):171-185, 1995.

[40] M. Mohri, F. Pereira and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, **16**(1):69-88,2002.

[41] N. Morgan and H. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proc. IEEE,* **83**(5):742-770,1995.

[42] Y. Morlec, G. Bailly and V. Aubergé, "Synthesizing attitudes with global rhythmic and intonation contours," *Proc. Eurospeech*, 2:219-222, 1997.

[43] H. Ney and S. Ortmanns, "Progress in dynamic programming search for LVCSR," *Proc. IEEE,* **88**(8):1224-1240, 2000.

[44] A. Le *et al.,* "The 2002 NIST RT Evaluation Speech-to-Text Results," *Proc. RT02 Workshop,* 2002. `http://www.nist.gov/speech/tests/rt/rt2002/`

[45] D. O'Brien and A. Monaghan, "Concatenative Synthesis Based on a Harmonic Model," *IEEE Trans. Speech and Audio Processing*, **9**(1):11-20, 2001.

[46] M. Ostendorf, V. Digalakis and O. Kimball, "From HMMs to Segment Models: A Unified View of Stochastic Modeling for Speech Recognition," *IEEE Tran. on Speech and Audio Processing,* **4**(5):360-378, 1996.

[47] M. Ostendorf and H. Singer, "HMM Topology Design using Maximum Likelihood Successive State Splitting," *Computer Speech and Language,* **11**(1):17-42, 1997.

[48] B. Pellom, *Enhancement, Segmentation, and Synthesis of Speech with Application to Robust Speaker Recognition*, Duke University, Ph.D. Dissertation, Electrical Engineering, 1998.

[49] M. Plumpe, A. Acero, H. Hon and X. Huang, "HMM-based smoothing for concatenative speech synthesis," *Proc. ICSLP*, 6:2751-2754, 1998.

[50] D. Reynolds, "Automatic Speaker Recognition using Gaussian Mixture Models," *MIT Lincoln Laboratory Journal*, **8**(2):173-192, 1996.

[51] R. Rosenfeld, "Two decades of statistical language modeling: where do we go from here?," *Proc. IEEE,* **88**(8):1270-1278, 2000.

[52] K. Ross and M. Ostendorf, "Prediction of Abstract Prosodic Labels for Speech Synthesis," *Computer, Speech and Language,* **10**(3):155-185, 1996.

[53] K. Ross and M. Ostendorf, "A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis," *IEEE Trans. Speech and Audio Processing,* **7**(3):295-309, 1999.

[54] Y. Sagisaka *et al.,* "ATR ν-talk speech synthesis system," *Proc. ICSLP,* 1:483-486, 1992.

[55] M. Saraclar, H. Nock and S. Khudanpur, "Pronunciation modeling by sharing Gaussian desnitites across phonetic models," *Computer Speech and Language,* **14**(2):137-160, 2000.

[56] S. Schwarm and M. Ostendorf, "Text normalization with varied data sources for conversational speech language modeling," *Proc. ICASSP,* 1:789-792, 2002.

[57] I. Shafran and M. Ostendorf, "Use of higher level linguistic structure in acoustic modeling for speech recognition," *Proc. ICASSP,* 3:1643-1646, 2000.

[58] I. Shafran, M. Ostendorf, and R. Wright "Prosody and phonetic variability: Lessons learned from acoustic model clustering", *Proc. ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, 127-132, 2001.

[59] E. Shriberg and A. Stolcke, "Prosody modeling for automatic speech understanding: an overview of recent research at SRI," *Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, 13-16, 2001.

[60] K. Silverman, J. Bellegarda and K. Lenzo, "Smooth contour estimation in data-driven pitch modeling," *Proc. Eurospeech*, 2:1167-1170, 2001.

[61] R. Sproat *et al.*, "Normalization of Non-Standard Words," *Computer Speech and Language*, **15**(3):287-333, 2001.

[62] A. Stolcke, "Improvements to the SRI LVCSR system," presented at the NIST Rich Transcription Workshop, May 2002.

[63] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Trans. Speech and Audio Processing*, **9**(1):21-29, 2001.

[64] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. ICASSP*, 2:837-840, 2001.

[65] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilisitic transform for voice conversion," *IEEE Tran. Speech and Audio Processing*, **6**(2):131-142,1998.

[66] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," *Proc. ESCA/COCOSDA Workshop on Speech Synthesis*, 273-276, 1998.

[67] M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," *Proc. ICASSP*, 2:805-808, 2001.

[68] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *J. Acoust. Soc. Am.*, **107**(3):1697-1714, 2000.

[69] K. Tokuda et. al., "An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features," *Proc. Eurospeech*, 1:757-760, 1995.

[70] K. Tokuda et. al., "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," *Proc. ICASSP*, 1:229-232, 1999.

[71] K. Tokuda et. al., "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. ICASSP*, 3:1315-1318, 2000.

[72] J. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer, Speech and Language*, **8**(2):95-128, 1994.

[73] J. van Santen and R. Sproat, "High accuracy automatic segmentation," *Proc. Eurospeech*, 6:2809-2812, 1999.

[74] M. Wang and J. Hirschberg, "Automatic classification of intonational phrase boundaries," *Computer Speech and Language*, 6:175-196, 1992.

[75] F. Weber, L. Manganaro, B. Peskin and E. Shriberg, "Using prosodic and lexical information for speaker identification," *Proc. ICASSP*, 1:141-144, 2002.

[76] G. Webster, *Toward a psychologically and computationally adequate model of speech perception*, University of Washington, Ph.D. Dissertation, Linguistics, 2002.

[77] P. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer, Speech and Language*, **16**(1):25-48, 2002.

[78] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Trans. Speech and Audio Processing*, **9**(1):30-38, 2001.

[79] S. Young, "Large vocabulary continuous speech recognition," *IEEE Signal Processing Magazine*, **13**(5):45-57, 1996.

[80] S. Young and P. Woodland, "State clustering in HMM-based continuous speech recognition," *Computer, Speech and Language*, **8**(4):369-384, 1994.

[81] J. Yi and J. Glass, "Natural-sounding speech synthesis using variable-length units," *Proc. ICSLP*, 4:1167-1170, 1998.