FOREWORD

MODIFICATION OF SPEECH: TRIBUTE TO MIKE MACON

Jan van Santen

This Foreward provides an overview of, and puts in perspective, the contributions of Mike Macon to text to speech synthesis (TTS). The core of his work consists of signal-processing algorithms that modify speech. Major opportunities exist for TTS systems that modify prosody of acoustic units, eliminating the need to search for units with the required prosody. However, the challenges to make prosodic modification-based systems sound more natural are formidable. Macon hasmodification-based played a role in several projects aimed at these challenges.

Introduction

Mike Macon's work in TTS focused on *signal-processing algorithms that modify speech.* Besides these core interests, Macon contributed to a larger group of projects, all focused on TTS. We discuss these projects and put his work in perspective by analyzing the role of speech modification—with special emphasis on *prosodic* speech modification—in TTS.

Prosody in TTS Systems

Two Procedures for Generating Prosody

In most TTS systems, the computation of prosody starts with a text analysis step in which prosodic mark-up tags are computed from text. What matters for text analysis quality is not only accuracy but also the level of detail of the prosodic tags. For example, speakers use many different types of pitch accent, degrees of accent strength, and types of phrase boundary. Such vari-

ations convey subtle shades in meaning [1]. However, most text TTS analysis systems make only coarse distinctions (e.g., emphasized vs. not emphasized, or comma vs. period) and predict even these poorly.

The next step consists of rendering these tags and can be performed via two quite different procedures. In *prosodic modification-based methods, quantitative target values* are computed from prosodic tags, units are retrieved from a speech corpus, and the units are modified to attain target values and then concatenated.

In *unit selection-based methods*, a tagged speech corpus is searched for units with matching prosodic tags, and the units are concatenated, optionally with some smoothing.

These two procedures represent extreme corners of a cube whose dimensions are (1) whether target values are computed, (2) whether prosodic modification takes place, and (3) the variety of prosodic contexts of each phoneme sequence in the speech corpus.

Both methods face serious challenges. For prosodic modification based methods, the quality of the speech generated depends on the degree to which target values express the prosodic tags, the naturalness of the target values, the difference between the original values of the units and the target values, and the adequacy of the signal-modification methods.

For unit selection-based methods, the quality of a given utterance depends on the probability that units are available that have the proper prosody and can be connected without audible seams. The key challenges here are how to create training text (i.e., the text used for creating the speech corpus) that has adequate *coverage* of the target domain without excessive amounts of recordings [2] and how to obtain recordings that are highly consistent. In evaluating these approaches to TTS, it is important to distinguish TTS applications in terms of the *combinatorial complexity* of the domain, defined as the number of phoneme sequence/prosodic context combinations that can occur in the domain. If an application has an intermediate level of combinatorial complexity, if little or no mark-up control is required, and if footprint is not an issue, then unit selection-based systems are currently optimal. For lower levels of complexity, word-splicing systems are optimal. And for higher levels of complexity, particularly if mark-up control is required or footprint is an issue, prosodic modification-based systems are needed. Unfortunately, the size of this third category of applications appears to be the largest, yet current prosodic modification-based systems do not provide adequate quality. *This means that improved prosodic modification-based synthesis is the core challenge faced by current TTS research.*

Improving Prosodic Modification-Based TTS

We now discuss projects that Macon either conducted, initiated, or influenced and that are all focused on improving prosodic modification-based TTS.

Signal-Processing Aspects of Prosodic Modification

Pitch and Timing Modification

In his Ph.D. thesis [3], Macon developed a speech synthesis system based on the sinusoidal model (also see [4], [5]) and extended the system for singing voice synthesis [6]. At CSLU, Macon developed the *OGIresLPC* module [7], a signal-processing back-end for Festival [8] based on pitch-synchronous residual-LPC encoding of the speech signal. The module enables high-quality time and pitch modification of diphones or nonuniform units and has superior smoothing capabilities to reduce concatenation artifacts. It is available for noncommercial use from [9].

Modifying Spectral Structure

It is by now well known that prosodic factors affect more than pitch, duration, and amplitude [10], [11]. Despite these effects, the main emphasis of current pitch- and timing-modification techniques appears to be on changing the spectral structure as little as possible. Two key questions are raised. First, how can we model the changes in spectral structure brought about by prosodic control factors? Second, how can we create new pitch- and timing-modification techniques that mimic these effects on spectral structure? Of course, the term *pitch-modification technique* is fundamentally misguided: instead, we should be dealing with *prosodic modification techniques* that perform an integrated multidimensional modification involving timing, pitch, and spectral structure. Moreover, the manner in which this is done may differ sharply depending on the prosodic control factor involved. For example, changing a phrase-medial unstressed syllable into a phrase-final unstressed syllable may require different modifications than are required for changing it into a phrase-medial stressed syllable. It is also important to realize that the recordings that are used for analysis and training focus on prosodic factors and not on pitch or timing in isolation. For example, one can instruct a speaker to use a uniformly higher-pitched voice. However, this may not result in the spectral changes brought about when pitch is locally increased as a result of a prosodic control factor; in fact, these recordings may be relevant more for singing than for speech.

Initial results show that spectral balance, as measured by the energy in broad frequency bands, can be predicted from prosodic control factors [12]. Currently, work is underway to control the spectral balance of output speech by applying a spectral weighting function to the amplitude parameters of the sinusoidal model.

Modifying Formant Trajectories

In most diphone-based systems, acoustic units are prosodic contex-independent. Phonemes approach invariant acoustic targets to allow for smooth concatenations between diphones. The result is that diphone speech often sounds overarticulated.

Macon and Wouters studied the effects of linguistic prosodic factors on the *rate-of-change* of formants in vowel and liquid transitions [13]. The prosodic factors that were investigated included lexical stress, pitch accent, word position, and speaking style. The results showed that the formant transitions were steeper in linguistically more prominent segments, that is, in stressed syllables, in accented words, in sentence-medial words, and in hyperarticulated speech. A numerical model was developed to predict changes in the formant rate-of-change based on the prosodic context of a transition.

The results of this study were integrated in a speech-modification algorithm to control the vowel quality of acoustic units during synthesis [14]. The method is based on predicting the desired formant rate-of-change of a speech unit based on the target prosodic context and the original prosodic context. For example, if a unit was recorded in a stressed, sentence-medial context but is to be synthesized in an unstressed, sentence-final context, the formant rate-of-change of the unit should decrease by a certain percentage. Modification of the actual formant rate-of-change is achieved by representing concatenated speech units using line spectral frequency (LSF) parameter trajectories and computing new trajectories that remain close to the original trajectories but also have the desired rate-of-change. Finally, speech is generated using the sinusoidal + all-pole signal representation, which allows preserving the original speech quality while modifying the formant structure.

Listening tests showed that the proposed technique enables modification of the degree of articulation of acoustic units with little degradation in the speech quality, and improves the naturalness of the synthesized speech.

Modifying Spectral Structure: Spectral Smoothing

Wouters and Macon invented a fusion unit-based smoothing technique [15], in which spectral information from two sequences of units are combined.

Concatenation units define initial spectral trajectories for the target utterance, and *fusion units* define desired transitions between concatenation units. The method uses a synthesis algorithm based on sinusoidal + all-pole synthesis of speech. Perceptual experiments showed that the method is highly successful in removing concatenation artifacts.

Perceptually Accurate Cost Measures

Recently, several studies have appeared that attempt to predict the quality of synthetic speech based on objective cost functions, including a study by Wouters and Macon [16], [17], [18]. Cost functions are important for a variety of reasons. First, unit selection-based methods need cost functions to select the optimal unit sequence. Second, prosodic modification-based systems need cost functions to preselect the best unit token for each unit type.

Generally, these studies focused on predicting audible spectral discontinuities from acoustic distance measures applied to the final and initial frames of the units. So far, this procedure has met with limited success. This is no surprise, because constructing a perceptually accurate cost function is challenging for a number of reasons. First, we cannot predict from these local acoustic costs whether the speech fragment generated by concatenation will have a natural trajectory. The challenge is to construct perceptually valid trajectory-based cost functions.

Second, unless a TTS system performs concatenation without any form of signal modification, the cost function must take into account the details of the combined concatenation and signal-modification operations. For example, in certain TTS systems, vowel portions of units are lengthened not by a uniform stretching operation but by inserting a linear trajectory between the two units. This can produce a natural-looking trajectory even when there is a spectral mismatch between the two units, provided that the directions of movement of the two units are compatible.

Third, any prosodic modification technique, whether applied to a small diphone inventory or to a large speech corpus, causes a certain level of quality degradation. An important question is how to predict this quality degradation as a function of the difference between the original and target prosodic contours. For example, should these differences be measured only in terms of F0? If so, should we measure these differences only on a frame-by-frame basis, or should we take into account differences in the direction of pitch change? Clearly, we need to conduct perceptual experiments to determine what types of prosodic differences between original and target values are easy and which are difficult to bridge.

Quick Adaptation to New Voices

Custom voices are desirable but expensive, because current technology requires a complete corpus to be recorded for each new voice. A technology that may change this is *voice transformation technology*. Kain and Macon [19] created a voice transformation method that captured features of the target speaker by using target speaker residuals, as follows. A baseline transformation system was constructed that transformed the spectral envelope as represented by the LPC spectrum, using a harmonic sinusoidal model for analysis and synthesis. The transformation function was implemented as a regressive, joint-density, Gaussian mixture model, trained on aligned LSF vectors by an expectation-maximization algorithm. The key innovation was the addition of a residual prediction module, which predicts target LPC residuals from transformed LPC spectral envelopes, using a classifier and residual codebooks. In a series of perceptual experiments, the new transformation system was found to generate transformed speech more similar to the target speaker than that generally by the baseline method.

Conclusions

We have seen how Macon's work, which started with a narrow focus on sinusoidal modeling during his graduate student years, has contributed to a large array of projects on TTS.

His work was rewarded by the prestigious NSF Career Development Award as well as by several research grants and honors, such as serving on the Speech Technical Committee of the IEEE Signal Processing Society.

Those who worked closely with him were delighted by the sense he exuded that you can do anything as long as you try, focus, plan, and get cheerfully excited when meeting new obstacles.

Astonishingly, he displayed the same attitude when he heard about his initial diagnosis: He called and told the devastating news in the same way that someone else would report being stuck in traffic. Mike had a great and provocative sense of humor. But this hid his basically shy and intensely warm nature. He was one of those rare people whom no one forgets, even after meeting him just once.

Acknowledgements

The research reported in this paper was supported in part by funds from Intel Corporation, Nippon Telephone and Telegraph, and National Science Foundation grants 0082718 and 0205731.

Bibliography

[1] D. Ladd, *Intonational phonology.* Cambridge, UK: Cambridge University Press, 1996.

[2] J. van Santen, "Prosodic modeling in text-to-speech synthesis," in *Proc. of Eurospeech-97*, (Rhodes), September 1997.

[3] M. W. Macon, *Speech synthesis based on sinusoidal modeling.* PhD thesis, Georgia Tech., October 1996.

[4] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proc. of the International Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 361–364, May 1996.

[5] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 6, pp. 557–560, November 1997.

[6] M. Macon, L. Jensen-Link, J. Oliverio, M. Clements, and E. George, "A system for singing voice synthesis based on sinusoidal modeling," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, ICASSP97, 1997.

[7] M. Macon, A. Cronk, J. Wouters, and A. Kain, "Ogireslpc: Diphone synthesizer using residual-excited linear prediction," Tech. Rep. CSE-97-007, OGI, 1997.

[8] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *Third ESCA Workshop on Speech Synthesis*, (Jenolan Caves, Australia), 1998.

[9] "Ogireslpc." [Online] Available: http://cslu.cse.ogi.edu/research/tts.htm.

[10] A. Sluijter, *Phonetic correlates of stress and accent.* Holland Institute of Generative Linguistics, 1995.

[11] G. Fant, A. Kruckenberg, S. Hertegard, and J. Liljencrants, "Accentuation and subglottal pressure in Swedish," in *Proc. ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, (Athens), September 1997.

[12] J. van Santen and X. Niu, "Prediction and synthesis of prosodic effects on spectral balance," in *Workshop on Speech Synthesis*, (Santa Monica, California), IEEE, 2001.

[13] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. I. Analysis," *Journal of the Acoustical Society of America*, 111(1):417–427, 2002.

[14] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. II. Synthesis," *Journal of the Acoustical Society of America*, 111(1):428–438, 2002.

[15] J. Wouters and M. Macon, "Unit fusion for concatenative speech synthesis," in *Proceedings ICSLP*, (Beijing, China), 2000.

[16] J. Wouters and M. W. Macon, "A perceptual evaluation of distance measures for concatenative speech synthesis," in *Proc. of the International Conf. on Spoken Language Processing*, vol. 6, pp. 2747–2750, November 1998.

[17] E. Klabbers and R. Veldhuis, "On the reduction of concatenation artifacts in diphone synthesis," in *Proc. ICSLP*, (Sydney, Australia), 1998.

[18] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. of the 26th International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (Salt Lake City, Utah), pp. 837–840, 2001.

[19] A. Kain and M. Macon, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, ICASSP01, 2001.