MODIFICATION OF SPEECH: A TRIBUTE TO MIKE MACON

Jan van Santen, Johan Wouters, and Alexander Kain

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, OR 97006, USA

ABSTRACT

This paper 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. The paper argues that major opportunities exist for TTS systems that modify prosody of acoustic units, instead of searching for units having the required prosody. However, the challenges to make prosodic modification based systems sound more natural are formidable. The paper provides an overview of the several projects aimed at these challenges, and in which Macon played a role.

1. INTRODUCTION

Mike Macon's work in text-to-speech synthesis (*TTS*) focused on *signal processing algorithms that modify speech* These algorithms fall into three classes: *Prosodic modification*, *Singing voice synthesis*, and *Voice transformation.* Besides these core interests, Macon also contributed to a larger group of projects, all focusing on prosody. This paper discusses these projects, and puts his work in perspective by analyzing the role of speech modification – with special emphasis on *prosodic* speech modification – in TTS.

2. PROSODY IN TTS SYSTEMS

For purposes of TTS, we distinguish between two meanings of the term "prosody". *(i) Linguistic prosodic control factors* are factors computed from text, such as word stress, sentence accent, proximity to phrase boundaries, or any factor other than phoneme identities. *(ii) Acoustic prosodic features* are the acoustic correlates of these factors, and include the standard features of pitch, amplitude, and timing.

2.1. 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. E.g., speakers use many different types of pitch accent, degrees of accent strength, and type of phrase boundary. This serves to 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.

(i) Prosodic modification based methods.

1. Compute *quantitative target values* from prosodic tags. Examples of target values include durations of phonetic segments, and per-frame quantities such as F_0 and amplitude.
2. Access speech corpus and retrieve units. The corpus typically consists of diphones recorded in a fixed phonemic and prosodic context.
3. Modify units to attain target values, and concatenate.

(ii) Corpus based methods.

1. Search tagged speech corpus for units with matching prosodic tags. The corpus may contain hours of speech, based on a variety of text types. At times, a diphone corpus is included in the larger corpus.
2. Concatenate units, optionally with some smoothing.

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

2.2. Key challenges in prosody generation

2.2.1. *Challenges for prosodic modification based methods*

The quality of the speech generated by modification based methods depends on:

This material is based upon work supported by the National Science Foundation under Grants No. 0117911 and 0082718, and by Intel and NTT. Wouters is now at SVOX AG.

0-7803-7395-2/02/$17.00 ©2002 IEEE

1. Naturalness of target values. A visual comparison of synthetic and natural F_0 contours shows, for most TTS systems, striking differences. Natural pitch contours can be described as smooth underlying curves that are interrupted and perturbed by segmental effects and voice fluctuations (creaking, jitter). Synthetic contours typically lack these perturbations, and also have simpler shapes than the underlying smooth curves in natural pitch; for example, they may be triangular, have linear segments interconnecting "pitch targets", or have a flat-hat contour. None of these shapes are found in natural speech. It has been argued that such stylized pitch contours cannot be distinguished from natural pitch contours, but many of these experiments were carried out with low-quality TTS systems and now should be replicated with the best current TTS technology.
2. Degree to which target values express the prosodic tags. The prosodic quality of the generated speech depends not only on whether the synthetic contours mimic contours found in natural speech, but also on whether the generated pitch contours reflect all distinctions made by the tags. For example, there may be 5 levels of emphasis tags, but this will find expression in the generated speech only if the target values preserve these distinctions.
3. Difference between original values of the units and target values. Signal modification methods work better if only small differences have to be bridged.
4. Adequacy of signal modification methods. Signal modification methods that do not preserve details of the recorded speech, such as LPC based synthesis with synthetic excitation, introduce fewer artifacts than methods that do preserve original detail, such as various PSOLA based methods. Of course, the problem with the former is that *all* speech has a compromised voice quality, regardless of the amount of prosodic modification.

2.2.2. Corpus based methods

Corpus based methods have the problem that the space of units (i.e., combinations of phoneme sequences [up to a certain length, e.g. 5; or of a certain type, e.g. syllables or words] and prosodic contexts) is not only quite large but also has a distributional feature whereby the chances that an arbitrary input sentence requires a rare unit is near certainty: The distribution has the *Large Number of Rare Events* (*LNRE*) property [2, 3, 4, 5]. In addition, as shown in [3], there are systematic distributional differences between text genres or applications: units that occur frequently in one application may not occur at all either in another application or in a general-purpose corpus such as on-line newspaper text. Finally, even if units locally possesses the proper prosody, there is no guarantee that the acoustic prosody that results after concatenation will be proper. The reason is that prosodic tags under-determine acoustic features. There might be F_0 mismatches, or the resulting overall pitch contour shape may not form a smooth pattern with continuous first and second derivatives. And indeed, it is fair to say that currently even the best corpus based TTS systems have precisely these problems.

3. APPLICATION DIMENSIONS

A few decades ago, rule based systems such as MITalk [6] were dominant. Since the mid-eighties, diphone based systems such as the Bell Labs TTS system [7] became prominent, and the mid-nineties saw the rise of corpus based systems such as CHATR [8] and NextGen [9]. Currently, these three broad categories co-exist. In addition, within these classes there is a wide variety of systems. As a result, current TTS systems display a great diversity of approaches, each with its own strengths and weaknesses.

This diversity reflects not only the waxing and waining of TTS groups at key research laboratories, but also the diversity of TTS applications. In this section, we discuss some of the main dimensions on which applications vary. These dimensions are: (i) Domain size; (ii) Hardware limitations; and (iii) Importance of synthesis mark-up languages.

3.1. Domain Size

There exist numerous examples of excellent quality that can be obtained with *Word and Phrase Splicing* (*WAPS*) systems. WAPS systems have been successfully deployed whenever (i) the domain has a fixed vocabulary (which can be quite large), and (ii) the grammar is restrictive.

Whether it makes sense to use TTS depends on how its quality compares to that of WAPS and on the relative efforts involved in deployment. General-purpose TTS systems are not competitive with WAPS systems in those applications where the latter can be deployed, because there are several companies that provide WAPS systems at low cost, and the quality of these systems is generally higher, and at times much higher, than than that of general-purpose TTS systems. Recently, however, companies have started to provide *application-specific corpus based TTS systems.* These systems have application-specific speech corpora, pronunciation dictionaries, and additional system components. They can handle applications whose combinatorial complexity is too large for WAPS systems, but not too large for the LNRE problem to cause system failure. The key question about these systems is a commercial one: are there enough applications in this relatively narrow range on the combinatorial complexity scale?

3.2. Hardware limitations

Embedded devices are becoming increasingly more speech enabled as a result of dramatic increases in processing speed. However, there are still strict limits on the amount of "fast" memory; slow storage, which does not have such limitations, is currently too slow for real-time unit retrieval.

3.3. Importance of synthesis mark-up languages.

Synthesis mark-up languages such as VoiceXML [10] and SABLE [11] can play an important role in at least two ways. First, there are applications that have both fixed text and variable text; it is because of the latter that TTS systems can be considered instead of WAPS systems because proper markup can create high quality prosody. Manual mark-up can be used to optimize the rendition of the fixed text. Second, there are applications with domains whose restrictions are such that specialized pre-processors can be constructed that insert mark-up tags for optimal rendition. We note that the requirement that a system should be controllable with mark-up tags puts a premium on the quality of prosodic modification (for prosodic modification based systems), or on the availability of units corresponding to any constellation of these tags (for corpus based systems).

3.4. Summary of Application Space

These three dimensions cast light on which TTS approaches are optimal for a given application. 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 corpus based systems are currently optimal. For lower levels of complexity, WAPS based systems are optimal. And for higher levels of complexity, in particular 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.*

4. 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.

4.1. Signal Processing Aspects of Prosodic Modification

4.1.1. Pitch and timing modification

In his PhD thesis [12], Macon developed a speech synthesis system based on the sinusoidal model (also see [13, 14]), and extended the system for singing voice synthesis [15]. At CSLU, Macon developed the *OGIresLPC* module [16], a signal processing back-end for Festival [17], based on pitch-synchronous residual-LPC encoding of the speech signal. The module enables high-quality time and pitch modification of diphones or non-uniform units, and has superior smoothing capabilities to reduce concatenation artifacts. It is available for non-commercial use from [18].

4.1.2. Modifying Spectral Structure

Macon and van Santen (NSF 0082718) started a project on analysis and synthesis of spectral structure, based on the following considerations.

It is by now well-known that prosodic factors affect more than pitch, duration, and amplitude [19, 20, 21, 22]. 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 both 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 is 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 [23]. Currently, work is under way to control the spectral balance of output speech by applying a spectral weighting function to the amplitude parameters of the sinusoidal model.

4.1.3. Modifying Formant Trajectories

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

Macon and Wouters studied the effects of linguistic prosodic factors on the *rate-of-change* of formants in vowel and liquid transitions [24]. 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, i.e. in stressed syllables, in accented words, in sentence-medial words, and in hyper-articulated 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 [25]. 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 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 to preserve the original speech quality while modifying the formant structure.

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

4.1.4. Modifying Spectral Structure: Spectral Smoothing

Wouters and Macon invented a "fusion unit" based smoothing technique [26], 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.

4.2. 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 [27, 28, 29]. Cost functions are important for a variety of reasons. First, corpus based methods need cost functions to select the optimal unit sequence. Second, prosodic modification based systems need cost functions to pre-select 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, one 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 F_0? If so, should one only measure these differences on a frame-by-frame basis, or does one take into account differences in the direction of pitch change? Clearly, we need to conduct perceptual experiments to determining what types of prosodic differences between original and target values are easy and which ones are difficult to bridge.

4.3. Corpus design

Corpus design is critical for the quality of corpus based TTS. There are many ways in which a corpus can be designed. At the simplest level, randomly selected text is used. At a more sophisticated level, text is selected having good diphone coverage, using greedy methods [30]. Greedy selection can also be applied to units other than diphones, such as words, prosodically tagged N-phones, or syllables. It is also recommended to use a heterogeneous corpus consisting of specialized sub-corpora, e.g. for diphones, number sequences, names, phrases from frequently used applications and dates. Application-specific corpora fall in this third category (subsection 3.1).

If at least some degree of prosodic modification is used, new possibilities emerge, which will be discussed next.

As stated earlier, prosodic modification methods produce better results when only a small modification needs to be made. This makes it important that we understand (i) which types of prosodic changes are likely to produce quality degradation, and (ii) how to label the text corpus (from which to-be-recorded would be selected with greedy methods) with tags that are maximally predictive of prosodic effects.

To elaborate on the latter, consider two extremes. If a text corpus is labeled using only two binary tags, stressed vs. unstressed and phrase-final vs. phrase-medial, there would be too much acoustic variability within each of the four resulting prosodic classes. At the other extreme, extremely detailed mark-up is used (e.g., in terms of parts-of-speech in a window of five words, distance to the next comma, word length, etc.), then the combinatorial space would become too large to be covered by the to-be-recorded text. What is needed is a tagging scheme that only makes those distinctions that are necessary to guarantee that speech intervals with the same tags are similar in terms of prosodic features (or can be modified distortion-free to be similar.)

A paper presented at this conference [31] investigates whether the *foot* concept provides such a scheme. (A foot is defined as a sequence of syllables of which only the first carries a pitch accent.) This work is an outgrowth of earlier discussions with Macon and Vincent Pagel. The underlying idea is that accent-lending up-down pitch movements are associated with a foot, and that the local pitch values and directions can be predicted from the location of a syllable in a foot, foot length, and the position of the foot in a phrase [32]. The paper suggests using foot-tagged di- and triphones as basic units.

4.4. 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 [33] 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 the baseline method.

A different technique is suggested by a small footprint TTS method [34], in which diphones are reconstructed from a small set of acoustic "basis vectors" using an asynchronous non-linear interpolation operation. To create a new voice, only recordings are required to estimate these vectors.

5. CONCLUSIONS

We have seen how Macon's work, starting out with a narrow focus as a graduate student on sinusoidal modeling, has contributed to a large array or projects that all focus on some aspect of prosodic modeling in TTS. It is our hope that his contribution, and these projects, will also inspire others to focus their efforts on how to make prosodic modification based systems sound better.

6. REFERENCES

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

[2] J. van Santen, "Timing in text-to-speech systems," in *Proc. of Eurospeech-93*, Berlin, 1993, vol. 2, pp. 1397–1404.

[3] J. van Santen, "Combinatorial issues in text-to-speech synthesis," in *Proceedings of Eurospeech-97*, Rhodes, September 1997.

[4] B. Moebius, "Rare events and closed domains: Two delicate concepts in speech synthesis," in *Workshop on Speech Synthesis*, Pilochry, 2001, ESCA.

[5] H. Baayen, *Word Frequency Distributions*, Kluwer, Dordrecht, The Netherlands, 2000.

[6] J. Allen, S. Hunnicut, and D.H. Klatt, *From text to speech: The MITalk System*, Cambridge University Press, Cambridge, U.K., 1987.

[7] Richard Sproat, Ed., *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, Kluwer, Boston, MA, 1997.

[8] N. Campbell and A. Black, "CHATR: a multi-lingual speech re-sequencing synthesis system," in *Proc. of Institute of Electronic, Information and Communication Engineers-89*, Tokyo, 1996.

[9] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T NextGen system," in *Proceedings of the Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, March 15–19, 1999*, Berlin, Germany, 1999.

[10] "Speech synthesis markup requirements for voice markup languages," http://www.w3.org/TR/voice-tts-reqs/, 1999.

[11] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington, "SABLE: A standard for TTS markup," http://www.research.att.com/ﬁws/SABPAP/sabpap.htm, 1987.

[12] M. W. Macon, *Speech synthesis based on sinusoidal modeling*, Ph.D. thesis, Georgia Tech., October 1996.

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

[14] 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.

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

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

[17] P. Taylor, A. Black, and R. Caley, "The architecture of the Festival speech synthesis system," in *Third ESCAWorkshop on speech synthesis*, Jenolan Caves, Australia, 1998.

[18] "Ogireslpc," http://cslu.cse.ogi.edu/research/tts.htm.

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

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

[21] M. Swerts and R. Veldhuis, "Interactions between intonation and glottal-pulse characteristics," in *Proceedings ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, September 1997.

[22] N. Campbell and M. Beckman, "Stress, prominence, and spctral tilt," in *Proceedings ESCA Tutorial and Research Workshop on Intonation: Theory, Models and Applications*, Athens, September 1997.

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

[24] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. I. Analysis," *Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 417–427, 2002.

[25] J. Wouters and M. Macon, "Effects of prosodic factors on spectral dynamics. II. Synthesis," *Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 428–438, 2002.

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

[27] 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*, November 1998, vol. 6, pp. 2747–2750.

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

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

[30] J. van Santen, "Diagnostic perceptual experiments for text-to-speech system evaluation," in *Proceedings ICSLP '92*, 1992, pp. 555–558.

[31] E. Klabbers and van Santen, "Prosodic factors for predicting local pitch shape," in *Workshop on Speech Synthesis*, Santa Monica, California, 2001, IEEE.

[32] J. van Santen and J. Hirschberg, "Segmental effects on timing and height of pitch contours," in *Proceedings ICSLP '94*, 1994, pp. 719–722.

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

[34] A. Kain and J. van Santen, "Compression of acoustic inventories using asynchronous interpolation," in *Workshop on Speech Synthesis*, Santa Monica, California, 2001, IEEE.