Applications of Computer Generated Expressive Speech for Communication Disorders

Jan van Santen, Lois Black, Gilead Cohen, Alexander Kain, Esther Klabbers, Taniya Mishra, Jacques de Villiers, and Xiaochuan Niu

Center for Spoken Language Understanding
OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, Oregon 97006, USA
vansanten@ece.ogi.edu

Abstract

This paper focuses on generation of expressive speech, specifically speech displaying vocal affect. Generating speech with vocal affect is important for diagnosis, research, and remediation for children with autism and developmental language disorders. However, because vocal affect involves many acoustic factors working together in complex ways, it is unlikely that we will be able to generate compelling vocal affect with traditional diphone synthesis. Instead, methods are needed that preserve as much of the original signals as possible. We describe an approach to concatenative synthesis that attempts to combine the naturalness of unit selection based synthesis with the ability of diphone based synthesis to handle unrestricted input domains.

1. Introduction

Text-to-Speech synthesis (TTS) has made substantial progress over the past decade, with notable improvements in voice quality and text analysis. However, at a recent TTS workshop (*IEEE Workshop on Speech Synthesis*, Santa Monica, California, September 2002), a strong consensus was expressed that prosody was still a major weak point. And indeed, even the best systems still too often produce utterances with inappropriate prosody: The wrong words are emphasized, phrase boundaries are not appropriately indicated, and there is no prosodic structure for longer stretches of speech. As a result, comprehension is difficult and the overall listening experience is disconcerting because of the impression that the speaker does not understand the text.

Prosody of current TTS systems is affectively neutral, and generally uses fairly "flat" intonation contours. If even this already constitutes a challenge, the bar is raised considerably higher when we demand that the prosody must be expressive. A question to ask is whether we are ready for this next big leap.

This paper discusses some of the challenges, and presents a case for why we should try to work on expressive speech synthesis: Expressive speech can provide useful diagnostic and remedial tools for certain communication disorders. We then discuss an approach that may be helpful in accomplishing this goal.

2. Affective Prosody in Speech Synthesis

2.1. Affective Prosody in Humans: General Features

Acoustic correlates of affect include pitch range, loudness, loudness range, overall speech rate, temporal pattern, and voice quality (e.g., [1, 2]). Although older experiments focused primarily on pitch range, loudness, and overall speech rate, there is strong evidence that affective prosody involves all of these factors (e.g., [3, 4]).

It is also known that perception of vocal affect, in experiments where no other cues are present is not terribly accurate (e.g., [5]). We speculate that this lack of accuracy has to do with the fact that, in natural situations, vocal affect is rarely perceived in a similarly context-free situation: Typically, there are numerous additional cues that provide information about the affective state of the speaker, including facial affect, contents of the utterance, preceding dialogue, preceding events, and the situation (e.g., a gunshot, a birthday cake); in addition, past experiences listening to the same speaker in various affective states may provide an acoustic frame of reference [5]. Although humans can be uncannily sensitive to subtle vocal affect cues, this ability might come into full play only under conditions where most of these other sources of information are also available. E.g., it may only be possible to note insincerity, or the precise type of anger, in a speaker's voice in the presence of other cues.

An illustration of the importance of non-vocal affect cues is provided a demonstration of what can be called a "Prosodic McGurk" effect. The process for generating these stimuli is straightforward. Audio-visual recordings are made of a sentence rendered with different types of affect. The audio portions of a pair of recordings representing different affects are lip-synched and re-combined with the video portion of the opposite recording. The process involves standard dynamic time warping and temporal modification algorithms. Interestingly, pilot results indicate that some individuals are far more likely than others to integrate conflicting audio- and video signals into new affective percepts.[1] Systematic studies are now under way that are specifically targeted on whether children with Autism are less likely to integrate these signals than children in the general population.

In conclusion, although the study of vocal affect by itself is important for many good reasons, the study of vocal affect might further gain in importance when vocal affect is observed in the context of the additional cues that are normally associated with affect.

2.2. Obstacles for the Synthesis of Expressive Speech

In the Introduction, we claimed that the generation of affectively neutral prosody has turned out to be a serious challenge, and that the generation of expressive prosody is even more difficult.

[1] See http://cslu.cse.ogi.edu/pub/vansanten/Public/EURO2003/.

We present here some reasons for this claim.

We use the term *unit selection based speech synthesis* for methods that search a labeled speech corpus for a sequence of speech intervals such that the sequence of associated labels matches a symbolic representation of the input text. These labels can be orthographic words or phrases ("phrase splicing") or arbitrary sequences of phoneme labels, and either type of labels can be further tagged with prosodic labels (e.g., "stressed".) The search process is followed by a concatenation process, in which the retrieved speech intervals are glued together without any further prosodic modification. Typical examples include CHATR [6] and NextGen [7].

A critically important feature of unit selection based speech synthesis relevant for the generation of vocal affect is that, exactly because no prosodic modification takes place, all acoustic correlates of affective prosody are preserved. Older, diphone-based concatenative systems (e.g., [8]) have a small set of diphone units that are not differentiated by prosodic features, let alone by affect tags. This means that for a given affect, target intonation, timing, and other acoustic parameters must be computed, and that subsequently these target parameters must be imposed on the diphones. To make this work, a deep understanding is needed of all acoustic correlates of affect and their interrelations. Thus far, this understanding is largely absent.

However, arguments presented elsewhere [9] about *combinatorial problems* inherent in unit selection based speech synthesis are relevant; in fact, the arguments apply with even more force when applied to expressive speech synthesis. In unrestricted domain synthesis (but also in seemingly restricted contexts such as synthesis of names in U.S. English or email reading), the number of combinations of phoneme sequences (e.g., up to length 4) and prosodic contexts (with enough detail to predict the local durational and pitch patterns) is quite large. Moreover, the probability mass of exceptionally rare combinations is sufficiently large that any given input text will almost certainly contain one of these combinations. This means that the corpus must contain a very large percentage of these exceptionally rare combinations, and hence be very large itself. Because of practical limitations, corpora cannot have this size. Typical flaws one observes in even the best current unit selection based speech synthesis systems reflect this problem: Quite often, the prosodic pattern is off – a word or syllable that should be emphasized is not, or a boundary tone that should be low is not. Some remedies have been invented, such as keeping the overall intonation flat so that prosodic errors are less audible, or including sub-corpora for specific application domains. But neither of these remedies address the fundamental problem, which is that the combinatorics of unrestricted language cannot be addressed with a method that lacks a combinatorial capability. Diphone-based concatenative systems have a combinatorial capability because they can combine any unit with any prosodic target pattern; however, these systems face a host of other problems due to the fact that too much information from the natural speech signal is lost and has to be re-created by rule.

In the light of these remarks, it is clear that generation of affectively expressive speech is particularly challenging for two reasons. One is that a further prosodic dimension is added, compounding the combinatorics. The second is that, by definition, one cannot use prosodically "flat" speech to hide inaccurate prosody.

In summary, both unit selection based speech synthesis and traditional diphone synthesis face serious challenges when the goal is that of generating compelling vocal affect in unrestricted domain synthesis.

3. Potential Value of Expressive Speech for Children with Language-Related Impairments.

Expressive speech synthesis provides special benefits for children with language-related disabilities. This is so for a number of reasons.

First, although far from being a universal feature of these children, poor expressive (i.e., output) prosody is a defining feature of certain subgroups, such as children with autism and pervasive developmental disorders (especially Asperger's Syndrome; [10]). Many of these children may have prosody that is either flat or stilted and unrelated to meaning.

Second, some children may have receptive (i.e., input) prosody impairments. This is the case for certain subgroups of children with developmental language disorders [11] and pervasive developmental disorders [12, 13]. Poor prosody contributes to weak non-verbal communication and processing of social cues, poor pragmatics, interactional difficulties, and emotional and behavioral problems.

Third, a common feature of therapeutic approaches is that they can involve the use of exaggerated prosody by the therapist. The assumption is that one needs exaggerated prosody to get through to the child; for example, to enhance the semantic meaning of what is being said, to heighten attention, or to affectively and socially engage the child [14, 15, 16]. "Motherese," a speaking style with exaggerated intonational contours, is known to play an important role in very early childhood for establishing the framework within which a child's joint attention and communicative competence grows [17]. The use of exaggerated prosody may be a significant aid to children who have inadequate comprehension and prosody.

Expressive TTS can play important roles for these children. First, if the intuitions underlying the therapeutic approaches discussed above are valid, then exaggerated prosody is a critical feature of language applications targeted at these children, even when the goal is not directly that of remediating expressive or receptive prosody in the child. Second, while we are not aware of attempts to teach prosody directly to these children, although it is a recommended component of treatment [18], methods have been developed for use with teaching English to non-native speakers [19] and the hearing impaired [20]. While systems for the latter group are necessarily based on visual representation of pitch contours, it is plausible that systems targeted at the former group of non-native speakers would be made more effective if TTS could be used that acoustically highlights critical intonational features. It seems also plausible that these methods can be adapted for use for children with language-related disabilities.

4. "Unit Quadruple" Based Synthesis

Unit selection based speech synthesis and traditional diphone synthesis represent extremes on a continuum. At CSLU, we are exploring methods that share some aspects with both methods while avoiding the pitfalls of either [21]. We describe here one such approach, currently under development, which involves corpora that have the property that *a small minority of units is recorded in all prosodic contexts and then used to generate prosodic contours for structurally similar units via prosody morphing methods.*

4.1. Unit Quadruples

Let P be a list of phone labels. Each label consists of an IPA phoneme label optionally combined with additional tags. Consider a set S of *units*, i.e. sequences of phones in P. (In standard diphone synthesis, S consists of a subset of all sequences consisting of two phone labels.) Any element x in S can be converted into a sequence x' consisting of corresponding phoneme class labels (nasals, voiced stops, etc.) S can be partitioned into a family of subsets, $S_1, S_2, \cdots$, such that within each subset all phone sequences have the same phoneme class labels. I.e., if x and y are in S_i for some i, then $x' = y'$ – the units are "structurally similar."

Let C be a complete list of all prosodic contexts. The system must be able to synthesize arbitrary combinations of S_i and C, but we can afford to make recordings of only a small subset of the product set $S_i \times C$. This can be accomplished by making selective recordings according to the principle illustrated in Table 1. For a given S_i, recordings are made for some unit u_i in all contexts in C, and for all units s in some context c_i. We call u_i and c_i the *shared unit* and *shared context* of S_i, respectively. As a result, for any unit u in S_i and context c in C we can find recordings of $< u_i, c_i >$, $< u, c_i >$, and $< u_i, c >$. The question is how to use these three recordings to generate the speech that completes the quadrangle, i.e. $< u, c >$.

		Prosody				
		.	c_i	.	c	.
	.	.	x	.	.	.
	u_i	x	[x]	x	[x]	x
Unit	.	.	x	.	.	.
	u	.	[x]	.	t	.
	.	.	x	.	.	.

Table 1: Unit Quadruple concept. For a given set of units, S_i, shared unit u_i and shared context c_i, with available recordings indicated by x, enable synthesis of target speech indicated by t. The specific recordings used for synthesis of t are boxed.

4.2. Time Warps

We now describe a procedure for generating $< u, c >$ from $< u_i, c_i >$, $< u_i, c >$, and $< u, c_i >$.

For arbitrary c_i and c, consider the time warp $w_1(t; u_i, c_i, u_i, c)$ [for short, w_1] computed using standard dynamic time warping based on e.g., the cepstral distance as cost [22, 23], and constrained such that it maps phone boundaries in $< u_i, c_i >$ onto the corresponding phone boundaries in $< u_i, c >$; typically, this added constraint is satisfied automatically because most phone boundaries are prominent in the cepstral domain.

This temporal relation associates time points in $< u_i, c_i >$ with time points in $< u_i, c >$ that are maximally similar in terms of spectral features. However, it does not necessarily associate time points that are prosodically similar. For this to occur, we must assume that, at the very least, the pitch movements have similar trends in the two recordings. For example, we do not want an increasing or up-down trend in one case, and a decreasing trend in the other case, because this has the potential of causing pitch modification distortions when we change the pitch contour in $< u, c_i >$ to the target pitch contour exemplified by $< u_i, c >$. It has been shown (e.g., [24]) that certain pitch contour modifications are far more harmful than others. One way to avoid pitch contour clashes is by including in the the labeling scheme (i.e., P) prosodic tags, e.g., "foot based tags" [25].

We now describe how to create $< u, c >$ by applying a time warp, w_3, to $< u, c_i >$. The principle used here is that *corresponding parts in u_i and u are stretched or compressed by the same percentages when context c_i is replaced by context c.* This can be formalized as follows. Let $w_2(t; u, c_i, u_i, c_i)$ [for short, w_2] denote the time warp that associates time points in $< u, c_i >$ with points in $< u_i, c_i >$. Let, for t in the time interval occupied by $< u_i, c_i >$, $slope_1(t)$ be the slope of w_1 at t, in other words, the local time stretching/compression factor. Then define:

$$w_3(t) = \Sigma_{\tau \leq t} slope_1[w_2(\tau)] \quad (1)$$

Thus, the temporal operation applied to $< u, c_i >$ in order to create $< u, c >$ consists of applying the time warp $w_3(t)$ to the frames in $< u, c_i >$. These frames may consist of line spectral frequencies, sinusoidal parameters, or other representations.

4.3. Prosody Transplant

Prosody transplant is performed by computing a *multiplicative transformation curve* characterizing the difference between $< u_i, c_i >$ and $< u_i, c >$, and then multiplying this curve with the curve in $< u, c_i >$, using appropriate time warping functions. In the F_0 domain, for t in the $< u, c_i >$ interval, the multiplicative transformation curve is given by:

$$Tran_{F_0}[t; u, c_i, u, c] = \frac{F_0[w_1\{w_2(t)\}; u_i, c]}{F_0[\{w_2(t)\}; u_i, c_i]} \quad (2)$$

where $F_0(t, x, y)$ is the F_0 value at time t for unit x in context y. The target F_0 curve is then given by multiplying $F_0(t, u, c_i)$ (optionally after first subtracting a local "phrase curve" – [26]) with (smoothed) $Tran_{F_0}[t; u, c_i, u, c]$ and then applying the time warp w_3.

Spectral balance parameters, which are known to be strongly associated with prosodic context [27, 28, 29, 30], can be captured by measuring the energy in formant-range frequency bands [31], and can be imposed on $< u, c >$ using Eq. 2.

5. Conclusions

We have described ongoing work in a research program focused on developing expressive speech synthesis. We believe that the challenge of producing expressive TTS is an important one for the TTS community for two reasons. First, it seems likely that fundamentally new methods must be developed. The Unit Quadruple approach is but one example of this. Second, research focused on non-traditional target populations, such as children with developmental language disorders, will run into unanticipated scientific obstacles; this is likely to stimulate new technological discoveries.

6. Acknowledgments

This research was supported in part by NSF Grant 0205731, to Jan van Santen, Alan Black, and Richard Sproat. The research was also supported by NSF Grant 0082718.

7. References

[1] S. J. L. Mozziconacci, "Prosody and emotions," in *Proceedings of Prosody 2002*, Aix-en-Provence, France, 2002.

[2] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[3] T. Johnstone and K. R. Scherer, "The effects of emotions on voice quality," in *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 3, San Francisco, 1999.

[4] C. Gobl, E. Bennett, and A. N. Chasaid, "Expressive synthesis: How crucial is voice quality?" in *Workshop on Speech Synthesis*. Santa Monica, California: IEEE, 2002.

[5] J.-A. Bachorowski, "Vocal expression and perception of emotion," *Current Directions in Psychological Science*, vol. 8, pp. 53–57, 1999.

[6] 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.

[7] 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.

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

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

[10] A. Klin, F. Volkmar, and S. Sparrow, *Asperger Syndrome*. New York, NY: Guilford Press, 2000.

[11] L. M. Black, "Subtypes of language disordered children at risk for social-emotional problems," Ph.D. dissertation, New School for Social Research University, New York, NY, 1989.

[12] A. Wetherby and B. Prizant, "Facilitating language and communication development in autism: Assessment and intervention guidelines," in *Autism: Identification, Education, and Treatment*, D. B. Zager, Ed. Hillsdale, NJ: Lawrence Erlbaum, 1999.

[13] C. Lord and R. Paul, "Language and communication in autism," in *Handbook of Autism and Pervasive Developmental Disorders (2nd edition)*, D. Cohen and F. Volkmar, Eds. New York, NY: John Wiley & Sons, Inc., 1997, pp. 195–222.

[14] S. Greenspan and S. Wieder, *The child with special needs*. Reading, MA: Addison-Wesley, 1998.

[15] B. Prizant, A. Wetherby, and P. J. Rydell, "Communication intervention issues for young children with autism spectrum disorders," in *Children with autism spectrum disorders: A developmental, transactional perspective*, A. Wetherby and B. Prizant, Eds. Baltimore, MD: Paul Brookes Publishing Company, 2000.

[16] C. Goossens, S. Crain, and Elder, *Engineering the Preschool environment for interactive, symbolic communication: An emphasis on the developmental period 18 months to five years*. Birmingham, AL: Southeast Augmentative Communication Conference Clinician Series, 1992.

[17] A. Fernald, "Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective," in *Language acquisition*, P. Bloom, Ed. Cambridge, MA: MIT Press, 1992, pp. 51–94.

[18] A. Klin and F. Volkmar, *Asperger's Syndrome: Guidelines for Treatment and Intervention*. New Haven, Connecticut: Yale Child Study Center, 1996, http://info.med.Yale.edu/chldstdy/autism.

[19] D. Hermes and G. Spaai, "A speech-melody trainer for foreign-language intonation," in *Proceedings of the Seventh Twente Workshop on Language Technology (TWLT7)*. Enschede, The Netherlands: University of Twente, 1994, pp. 115–116.

[20] G. Spaai, D. Derksen, D. Hermes, and P. Kaufholz, "Teaching intonation to young deaf children with the intonation meter," *Folia Phoniatrica et Logopedica*, vol. 48, pp. 22–34, 1996.

[21] J. van Santen, J. Wouters, and K. A., "Modification of speech: A tribute to mike macon," in *Workshop on Speech Synthesis*. Santa Monica, California: IEEE, 2002.

[22] J. van Santen, J. Coleman, and M. Randolph, "Effects of postvocalic voicing on the time course of vowels and diphthongs," *Journal of the Acoustical Society of America*, vol. 92, no. 4, Pt. 2, p. 2444, October 1992.

[23] J. van Santen, "Segmental duration and speech timing," in *Computing Prosody*, Y. Sagisaka, W. Campbell, and N. Higuchi, Eds. New York: Springer-Verlag, 1996.

[24] E. Klabber and J. van Santen, "Control and prediction of the impact of pitch modification on synthetic speech quality," in *Proceedings of Eurospeech-2003*, Geneve, Switzerland, September 2003.

[25] E. Klabbers and van Santen, "Prosodic factors for predicting local pitch shape," in *Workshop on Speech Synthesis*. Santa Monica, California: IEEE, 2002.

[26] J. van Santen, C. Shih, and B. Möbius, "Intonation," in *Multilingual Text-to-Speech Synthesis*, R. Sproat, Ed. Dordrecht, the Netherlands: Kluwer, 1997.

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

[28] 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.

[29] 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.

[30] 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.

[31] J. van Santen and X. Niu, "Prediction and synthesis of prosodic effects on spectral balance," in *Workshop on Speech Synthesis*. Santa Monica, California: IEEE, 2002.