ICSLP 2002 (In Press)

EXPRESSIVE SPEECH SYNTHESIS USING A CONCATENATIVE SYNTHESIZER

Murtaza Bulut[1*], *Shrikanth S. Narayanan*[*], *Ann K. Syrdal*[**]

[**]AT&T Labs-Research, Florham Park, NJ
[*]Department of Electrical Engineering and Integrated Media Systems Center
Speech Analysis and Interpretation Laboratory: http://sail.usc.edu
University of Southern California, Los Angeles, CA
mbulut@usc.edu, shri@sipi.usc.edu, syrdal@research.att.com

ABSTRACT

This paper describes an experiment in synthesizing four emotional states - anger, happiness, sadness and neutral – using a concatenative speech synthesizer. To achieve this, five emotionally (i.e., semantically) unbiased target sentences were prepared. Then, separate speech inventories, comprising the target diphones for each of the above emotions, were recorded. Using the 16 different combinations of prosody and inventory during synthesis resulted in 80 synthetic test sentences. The results were evaluated by conducting listening tests with 33 naïve listeners. Synthesized anger was recognized with 86.1% accuracy, sadness with 89.1%, happiness with 44.2%, and neutral emotion with 81.8% accuracy. According to our results, anger was classified as inventory dominant and sadness and neutral as prosody dominant. Results were not sufficient to make similar conclusions regarding happiness. The highest recognition accuracies were achieved for sentences synthesized by using prosody and diphone inventory belonging to the same emotion.

1. INTRODUCTION

It is usually rare to mistake synthetic speech for human speech. The complex nature of human speech, which comes from the fact that it varies depending on the speaking style and emotion of the speaker, makes it difficult to be imitated by synthetic speech. Intelligibility, naturalness and variability are three features used to compare synthetic speech with human speech [10]. In terms of intelligibility, a measure of how understandable speech is to humans, recent research has shown that synthetic speech can reach intelligibility levels of human speech. However, lack of variability, representing the changes in speech rate, voice quality, and naturalness, i.e. how "human" a synthesizer sounds, has impeded the general acceptance of synthetic speech, especially for extended listening. Incomplete knowledge of how factors such as the type of the material read, behavior of the audience, speaker's social standing, attitude and emotions, affect the speech signal has been a major problem in building more "human" sounding systems.

The need for human sounding text-to-speech synthesis (TTS) comes from the fact that it can greatly enhance applications based on human-machine interaction and make them simpler and more compelling. Imagine that a pleasant voice is reading your e-mails, web sites and books for you. When you have a question, you can ask a "virtual teacher" that adapts his voice depending on the topic and the nature of your responses and questions. Then, you can play games and watch films without realizing that you are hearing a synthetic voice.

Early attempts at imparting emotional quality to synthetic speech were based on rule-based TTS, including the pioneering efforts of Cahn [3]. The lack of naturalness in the speech synthesized using such schemes however poses a serious drawback. In this paper we describe the production of synthetic speech by concatenation of "emotional diphones" using Time-Domain Pitch Synchronous Overlap Addition (TD-PSOLA) [9] as the concatenation method. A similar approach has been applied to synthesize emotional speech in Spanish [8]. Listeners' recognition of emotion for Spanish showed that prosodic (supra-segmental) information alone was not enough to portray emotions. It was found that supra-segmental information characterized sadness and surprise while segmental components were dominant for cold anger and happiness. Studies on German emotional speech [6] showed that prosodic parameters, fundamental frequency and duration, were not enough to synthesize emotions. Increasing the parameter space by including voice quality parameters, spectral energy distribution, harmonics-to-noise ratio and articulatory precision has been shown to improve the recognition results for emotional Austrian German speech [13]. Experiments on synthesizing emotional speech using Japanese emotional corpora with CHATR [7] also support using an emotional inventory to synthesize emotional speech.

2. DATABASE COLLECTION

For the purpose of emotional speech synthesis reported in this paper, we chose to work with four target emotional states: anger, happiness, sadness, and neutral. First, we constructed five emotionally unbiased target sentences, i.e., sentences suitable to be uttered with any of the four emotions. The sentences we prepared were "I don't want to play anymore", "She said the story was a lie", "It was the chance of a lifetime", "They are talking about rain this weekend" and "OK, I'm coming with you".

Target diphones necessary to synthesize the five target sentences were determined. Next, four different source text scripts (one for each emotional state) were prepared for

[1] Most of M. Bulut's work for the paper was carried out at AT&T Labs.

recording; each source script included all the target diphones. Our aim in preparing the source text was to build emotionally biased sentences that could easily be uttered with the required emotion. The source sentences were declarative and on average 7 words long. The five target utterances were also included in each of the four inventories. In order to motivate and focus the speaker, each of the source sentences was accompanied by a one or two sentence scenario. These brief scenarios were prepared for eliciting happy, sad and angry inventories, and were not used for the neutral sentences. For example, "This is a wonderful life" and "I earned fifty million bucks" were scenario and source sentences, respectively, belonging to the happy inventory. Having such elicitation scenarios also helps to minimize the interpretation variations [11] that may result from speaker to speaker. It also increases the probability of getting the same effect from different speakers or from the same speaker at different times. This assumption is closely related to the cognitive emotion definition perspective that every emotion is associated with a particular appraisal [4].

In this paper we give results based on recordings obtained from a semi-professional female actress. A total of 357 source utterances (97 angry, 107 sad, 97 happy, 56 neutral) were recorded. The recordings were made in a sound-proof room at 48kHz sampling rate using a unidirectional, condensor, head-worn B&K capsule microphone. For synthesis, all files were later downsampled to 16kHz. The phonetic segmentation and alignment was first performed automatically with the Entropics' Aligner software that used a phonetic transcription dictionary prepared at AT&T Labs-Research. Labeling for all sentences was manually checked and corrected when necessary.

3. SYNTHESIS OF EMOTIONAL SENTENCES

In most studies, human emotions are categorized into basic distinct categories such as anger, happiness, sadness, fear, disgust, surprise and neutral. Although this approach is correct, especially from a pragmatic sense, it can be regarded as a somewhat gross representation because it ignores the variations that exist within each emotion. For example, both hot-anger (rage) and cold-anger (hostility) are treated under the same category, although they show different acoustical and psychological characteristics; similar examples can be provided for other emotions. The lack of a complete formal definition for each emotion and variations resulting from gender, personality and cultural differences (see Social Constructivist perspective [4] and [14]) makes it impossible to account for every small variation. Despite some disadvantages, interpreting a perceived emotion as one of the seven basic emotions has the major advantage of the Darwinian perspective [5], which holds that there are certain universal basic emotions, and all other emotions can be derived from them.

For this experiment we decided to test the possibility of producing basic emotions by mixing prosodic information and diphones belonging to distinct emotional states. Interpreting "set 1" as comprising prosodic information corresponding to angry, happy, sad and neutral target sentences and "set 2" as the diphone inventory for angry, happy, sad and neutral sentences, we produced 80 synthetic sentences by combining "set 1" and "set 2" properties for the five target sentences:

set 1 =Prosody of {angry, happy, sad, neutral} target sentences
set 2 =Inventory of {angry, happy, sad, neutral} sentences
set 1 x set 2 x {5 target sentences} = 80 synthetic sentences

For the synthesis of these 80 sentences, the Festival Speech Synthesis System [2] provided a simple method of diphone concatenation using an implementation of TD-PSOLA [9] that produces good quality synthetic speech with easy modification of pitch and duration and a low computational load. In the generation of the 80 synthetic test utterances, there were three basic steps: analysis, modification and concatenation. In the analysis step, the required prosodic (i.e. pitch and duration) information was calculated from the target sentences. The speech segments extracted from the source (inventory) sentences were then modified to match the prosodic target data and, finally, were concatenated. Diphones, selected manually, were the basic concatenation units used in this experiment.

4. RESULTS

Web-based listening tests were conducted with 33 adults who were unaware of the identity of the test stimuli. Fourteen participants (5 of 8 females and 9 of 25 males) were native speakers of English and 19 were nonnative. The listeners were allowed to play the test files as many times as they wished, and were asked to choose for each the most suitable emotion among angry, happy, sad and neutral options in a "forced choice" task. They also rated the success of expressing the emotion they had selected along a 5-point scale: excellent (5), good (4), fair (3), poor (2), bad (1). A total of 100 files, consisting of 80 synthetic sentences and 20 original (recorded) target sentences were presented in a random order that was different for each listener.

Figures 1 and 2 show, respectively, listeners' emotion recognition rates for the 20 original sentences and for 20 "matched" synthesized sentences in which both prosody and inventory were extracted from the same emotion. In the figures, the suffix "-L" marks listeners' choices; for example "Angry" indicates the intended emotion, while Angry-L represents what emotion listeners heard. Emotions are represented by their first initials, "p" indicates "prosody" and "i" indicates inventory; for instance ApAi represents the matched synthetic sentences that used Angry prosody information and the Angry inventory.

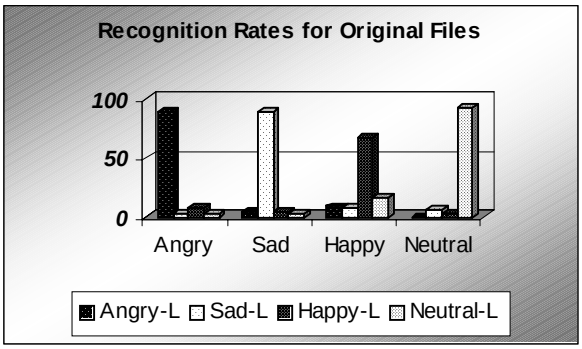

Figure 1: Recognition accuracy for natural target files: 89.1% Angry, 89.7% Sad, 67.3% Happy, 92.1% Neutral

Recognition rates for all original sentences (Figure 1) and for all matched synthetic sentences of each emotion (Figure 2) are significantly above the 25% chance level (as tested by one-sample t-Tests). Recognition accuracy for original and matched synthetic sentences was analyzed using a repeated measures ANOVA. There were significant differences in recognition accuracy among the four emotions: recognition rates observed for the Happy set were significantly lower than rates for the other three emotions in both the original and matched synthetic sets; either the speaker was relatively less successful in expressing happiness, or happiness is more difficult to recognize in isolated utterances. For the matched synthetic sentences, recognition accuracy for the Sad set was significantly higher than for the Neutral set. Happy and Neutral emotions were recognized more accurately in original sentences than in matched synthetic sentences, but Angry and Sad recognition rates were equivalent between original and matched synthetic versions.

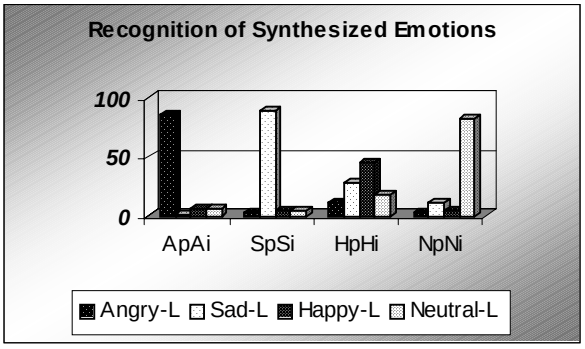

Figure 2: Recognition accuracy for synthesized emotions: 86.1% Angry, 89.1% Sad, 44.2% Happy, 81.8% Neutral "A", "S", "H", "N" denotes Anger, Sadness, Happiness and Neutral, respectively; "p" indicates prosody and "i" indicates inventory.

It is difficult to attribute the lower recognition of Happy and Neutral emotions in matched synthetic sentences than in the original sentences to the same cause. We can deduce that the lower recognition rate for synthetic happiness is in part due to the less successfully conveyed intended emotion of happiness in the Happy original target sentences and inventory. However, the same explanation does not hold for the Neutral emotion, the recognition accuracy for which in the original sentences was the highest of the four emotions.

Recognition rates and Average Success for the 16 different contour and inventory combinations are presented in Table 1. The two measures were significantly correlated (r = 0.60). Average Success was calculated by weighing excellent, good, fair, poor and bad responses by 5, 4, 3, 2 and 1, respectively.

According to Table 1, combinations of Angry inventory (Ai) with Angry prosody (Ap), Neutral prosody (Np) and Happy prosody (Hp) were classified, in most cases, as "angry", with the highest rate for the matched combination ApAi. Combinations of Sad prosody (Sp) with all other inventories were recognized as "sad" with an average of 80.6% accuracy. Synthetic sentences produced by employing Neutral prosody (Np) and Neutral, Happy and Sad inventories (Ni, Hi, Si) were recognized as "neutral", while NpAi was recognized as "angry" the majority of the time. Results for "happiness" show that it was mostly mistaken with "sadness" or a "neutral" emotional state. We observe that most successful recognition results were achieved for matched synthetic sets, i.e. when inventory and contour belonging to the same emotion were used together. It is also interesting to note that the combination of Neutral prosody and Angry inventory (NpAi) was recognized mostly as "anger", the combination of Sad prosody and Neutral inventory (SpNi); and Happy prosody and Neutral inventory (HpNi) as "sadness".

Pros. - Inv. Combination	Recognition Rate – Average Success			
	Angry-L	*Sad-L*	*Happy-L*	*Neutral-L*
ApAi	**86.1 - 4.1**	1.2 - 3.0	6.1 - 3.1	6.7 - 2.7
NpAi	**63.0 - 3.7**	3.6 - 3.2	1.2 - 3.0	32.1 - 3.2
HpAi	**59.4 - 3.4**	15.8 - 2.7	11.5 - 2.7	13.3 - 2.7
SpSi	2.4 - 3.3	**89.1 - 3.7**	4.8 - 2.6	3.6 - 2.8
SpNi	0.0 - 0.0	**89.1 - 3.6**	6.7 - 2.7	4.2 - 2.3
SpHi	1.8 - 3.0	**82.4 - 3.2**	11.5 - 3.3	4.2 - 2.1
SpAi	28.5 - 3.3	**61.8 - 3.2**	3.0 - 2.4	6.7 - 2.3
HpSi	15.2 - 3.3	**46.7 - 3.3**	23.6 - 3.2	14.5 - 3.1
ApSi	32.1 - 3.2	**37.6 - 3.0**	7.9 - 2.8	22.4 - 2.8
ApNi	15.2 - 2.9	**35.8 - 2.7**	17.0 - 2.8	32.1 - 3.0
HpNi	7.3 - 3.5	**35.2 - 3.2**	34.6 - 3.2	24.2 - 3.2
HpHi	10.3 - 2.9	27.3 - 3.0	**44.2 - 3.0**	18.2 - 3.1
ApHi	20.6 - 3.0	25.5 - 3.1	**29.7* - 3.2**	24.2 - 3.0
NpNi	3.0 - 3.2	10.9 - 3.3	4.2 - 3.0	**81.8 - 3.5**
NpHi	10.3 - 2.8	9.7 - 2.7	8.5 - 2.9	**71.5 - 3.3**
NpSi	13.3 - 3.5	17.6 - 3.1	2.4 - 3.7	**63.7 - 3.2**

Table 1: Recognition rates in percent and Average Success ratings (5=excellent and 1=bad) for the 16 possible prosody and inventory combinations. (* ApHi rate is not above chance)

Results based on the gender and language background of listeners are illustrated in Figure 3. Although the experiment was not designed to study systematically possible group differences, ANOVA results yielded no significant gender or language effects on recognition accuracy of either the original or the matched synthetic sentences.

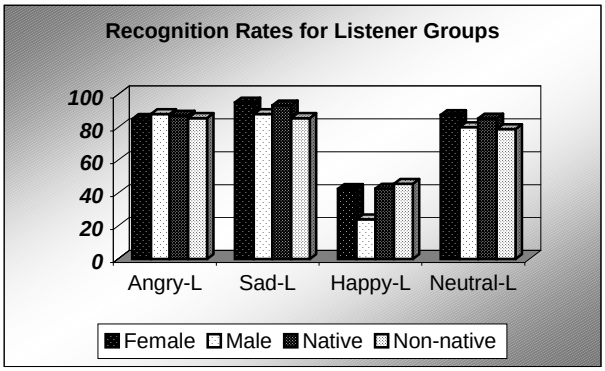

Figure 3: Recognition rates observed for matched synthetic sentences of each emotion for female, male, native and non-native listeners. There were no significant group differences.

5. DISCUSSION

The recognition results presented in Figure 2 and Table 1 show that anger, sadness and happiness can be synthesized fairly successfully by applying concatenative synthesis techniques. Recognition rates achieved for anger and sadness suggest that they were easier to synthesize when compared to happiness. Although, not directly comparable because of differences in the set of emotions and languages used, these results agree with those presented by other researchers [1, 6, 7, 8].

As seen in Table 1, taking the inventory of one emotion and mixing it with prosodic information of a different emotion gave lower recognition results than when the inventory and contour combination belonged to the same emotion. Based on these results, the ideal way to portray a particular basic emotion appears to be to use a separate database and separate prosodic models for each emotion.

We also observe that ApAi, NpAi and HpAi combinations were recognized as "anger". This suggests that segmental components (which include vocal quality and phonetic characteristics) were dominant in synthesizing anger. From the recognition results for SpNi, SpSi, SpHi and SpAi, we conclude that supra-segmental information determined "sadness". Listener ambiguity in the recognition of "happiness" prevented us from drawing similar conclusions. These results generally agree with experiments on synthesizing emotions by mixing diphones and prosody for the Spanish language [8] where anger and happiness have been classified as segmental emotions, and sadness as a prosodic emotion.

It is also seen (Table 1) that for most mismatched prosody and inventory combinations, the two emotions recognized with highest accuracy were the two used in the combination in question. For example, for NpAi, "anger" and "neutral", and for SpHi, "sadness" and "happiness" were the most frequently recognized emotions. This shows that both prosody and inventory are important in conveying emotions.

The most common feedback given by our listening test subjects was that some sentences conveyed different flavors of emotions than the ones listed as choices. This kind of listener feedback is promising and exciting because it is consistent with the Darwinian approach [4, 5] that all emotions can be derived from basic emotions. Future listening tests where listeners will be given an opportunity to choose among a larger set of emotions will be helpful in validating this hypothesis. A better understanding of this issue may reduce the need to record a separate inventory for each derived emotion.

The difficulty in expressing happiness for both original and synthesized sentences indicates the need for new experimental approaches. In addition to the text scenarios, use of visual aids such as pictures, videos, sounds, may help the actor/actress to express the required emotion more successfully. Since synthetic utterances depend on inventory, it is hoped such techniques will improve the artificial expression of happiness.

6. SUMMARY AND CONCLUSIONS

Demand for more "human-sounding" speech synthesis has created the need to synthesize emotions. In this paper we show that by using separately recorded inventories for anger, happiness, sadness and neutral emotions, and basic diphone concatenation synthesis with TD-PSOLA within the Festival System, some of these synthesized emotions can be reliably recognized by listeners. The recognition rate for anger was 86.1% with 4.1 Average Success rating (max = 5), for sadness, 89.1% with 3.7, for neutral emotion, 81.8% with 3.5, and for happiness, 44.2% with 3.0. Happiness was the most difficult emotion to convey with either natural or synthetic speech.

Segmental information was dominant in conveying anger, while prosody best characterized sadness and neutral emotion. Different combinations of inventory and prosody of basic emotions may provide synthesis of various intermediate emotional nuances. This is a topic of future research.

7. REFERENCES

[1] Abadjieva, E., Murray, I.R., and Arnott, J.L., "Applying Analysis of Human Emotional Speech to Enhance Synthetic Speech", *Proc. of Eurospeech*, pp. 909-912, Berlin, Germany, September 1993.

[2] Black, A., and Taylor, P., "The Festival Speech Synthesis System: System Documentation", Technical Report HCRC/TR-83. Human Communications Research Centre, University of Edinburgh, Scotland, UK, January 1997.

[3] Cahn, J.E., "Generating Expressions in Synthesized Speech", Master's Thesis, MIT, 1989. http://www.media.mit.edu/~cahn/masters-thesis.html

[4] Cornelius, R. R., "Theoretical Approaches to Emotion", *Proc. of ISCA Workshop on Speech and Emotion*, Belfast, September 2000.

[5] Ekman, P., Friesen, W., Sullivean, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Pitcairn, T., Ricci-Bitti, P., Scherer, K., and Tomita, M., "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion", *Journal of Personality and Social Psychology*, 53, 712-717, 1987.

[6] Heuft, B., Portele, T., and Rauth, M., "Emotions in Time Domain Synthesis", *Proc. of ICSLP*, Philadelphia, USA, October 1996.

[7] Iida, A., Campbell, N., Iga, S., Higuchi, F. and Yasumura, M., "A Speech Synthesis System for Assisting Communication", *ISCA Workshop on Speech and* Emotion, pp. 167-172, Belfast 2000.

[8] Montero, J.M., Arriola, G.J., Colas, J., Enriquez, E., and Pardo, J.M., "Analysis and Modeling of Emotional Speech in Spanish", *Proc. of ICPhS*, vol. 2, pp. 957-960, San Francisco, USA, 1999.

[9] Moulines, E., and Charpentier, F., "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", *Speech Communications*, vol. 9, pp. 453-467, December 1990.

[10] Murray, I.R., Arnott, J.L., and Rohwer, E.A., "Emotional Stress in Synthetic Speech: Progress and Future Directions", *Speech Communication*, Vol. 20, pp. 3-12 November 1996.

[11] Pell, M.D., "Influence of Emotion and Focus on Prosody in Matched Statements and Questions, *JASA*, 2001.

[12] Picard, R., "Affective Computing", *The MIT Press*, 1997.

[13] Rank, E., and Pirker, H., "Generating Emotional Speech with a Concatenative Synthesizer", *Proc. of ICSLP*, pp. 671-674, Sydney, Australia, 1998.

[14] Scherer, K.R., "A Cross Cultural Investigation of Emotion Inferences from Voice and Speech: Implications for Speech Technology", *Proc. of ICSLP*, Beijing, China, 2000.