Expressive Speech Synthesis and Modeling

1. Research Team

Project Leader:	Prof. Shrikanth Narayanan, *Electrical Engineering*
Graduate Students:	Murtaza Bulut
Undergraduate Students:	Ryan Cunningham

2. Statement of Project Goals

As human beings we communicate with each other through our feelings, which are expressions shaped by the experience and knowledge we have. Since every single state of humans can be related to a particular emotion, the role of emotions in communication cannot be underestimated. There have been studies of the human brain showing the impossibility to make appropriate decisions when the emotion-controlling centers in the brain are damaged, even if logical reasoning skills are intact [6]. Reasonably, the incorporation of emotions in the applications and systems trying to model human-like behaviors and effects is necessary and important.

Intelligibility, variability and naturalness are the features of speech that have been mostly used to measure the performance of synthetic speech [1]. In terms of intelligibility, the main focus in speech synthesis for many years, most speech synthesizers can now produce acceptable results. Unfortunately, the same is not true for the other two parameters: *Variability* and *Naturalness*. Variability is defined as the ability to change the characteristics of the voice depending on what is synthesized, while naturalness is - not very clearly and formally defined- a measure of how human-like the artificial speech sounds [1]. Only when the technology reaches a level where these three criteria are satisfied can speech synthesizers sounding like human be developed.

The main goal of the project is to contribute to the research of adding expressiveness to the synthetic speech. We aim to build a limited domain text-to-speech synthesizer that will be able to imitate the emotional characteristics of human speech. In order to achieve the desired naturalness the concatenative speech synthesis technology, which is based on concatenation of pre-recorded speech segments, will be employed in the project. Variability will be achieved by building a large database of emotional speech for different speakers and by employing data-driven transform functions that will enable us to change one emotion into another. In fact, an ultimate goal is to allow transformation of any emotional speech category to any other desired category, including the ability for synthesizing various intermediate expressive flavors in the spectrum of human emotions.

Initially, our work will be concentrated on five basic emotions: anger, sadness, happiness, frustration plus neutral.

The main goals of the project can be summarized as follows: (1) To build a database of emotional speech that can be used for other research purposes, (2) to develop transformation algorithms that will enable conversion of one emotional category into another, and finally using all these (3) to build a speech synthesizer that will model the acoustical changes in emotional human speech.

3. Project Role in Support of IMSC Strategic Plan

Enabling natural interactions with machines and other people, regardless of the space and time constraints is part of IMSC's vision. Spoken interactions are an important aspect of human communication. Research in expressive speech synthesis fits within the realm of enabling natural human machine interactions.

4. Discussion of Methodology Used

Firstly the database of emotional sentences was prepared. The database consisted of target sentences that we want to synthesize and of the supplementary sentences. The target sentences were chosen so that they will be uttered with one of the targeted emotions. On contrary the supplementary sentences were built as emotion specific sentences that contain all of the diaphones necessary to synthesize the target sentences. Five sets of sentences, one for each emotion (anger, happiness, sadness, frustration, neutral), were prepared for recording. Averaging the number of sentences in each set to 70 and adding also the target sentences, this means that approximately 400 sentences would be recorded per speaker. This amount of data will be very useful in characterizing and analyzing human speech in terms of its acoustical correlates. For the recording both non-experienced and experienced speakers will be used. The data obtained from non-experienced speakers will serve as a good comparison set to the "exaggerated" emotional data obtained from experienced (actors) speakers [2].

In order to produce the required synthetic output University of Edinburgh's Festival Speech Synthesis System [3] is used. Taking the prosody and duration information from the recorded target sentences TD-PSOLA [4] method has been used as a synthesis method. In addition, the information that will be obtained from the analysis of the recorded sentences, regarding the dependency of fundamental frequency and duration of the corresponding speech to each emotion units will be incorporated in the system. The fact that diaphone concatenation systems are not as robust as format synthesizers in terms of parameterization is the main difficulty in trying to get expressive synthetic output. In most concatenative synthesizers only F0, duration and in some cases intensity are controllable [5]. This is in fact the price we pay, to get more natural voice.

For transformation algorithms we concentrate on spectral conversion based on Gaussian mixture models (GMMs) for transforming diaphone units of one emotion category to another. The method, motivated by applications in voice conversion [7,8], is based on building GMM for joint spectral vectors, which are the combination of source and target vectors.

5. Short Description of Achievements in Previous Years

An initial database was created for 2 speakers (one male and one female). A detailed survey of the state of the art in the field was obtained. A detailed database was designed. Contributions were made to synthesizing "command" speech for a military application.

5a. Detail of Accomplishments During the Past Year

Experiment in synthesizing four emotional states - anger, happiness, sadness and neutral – using a concatenative speech synthesizer was performed. To achieve this, 5 emotionally unbiased (i.e., semantically) target sentences were prepared. Then, separate inventories, comprising the target diaphones, for each of the above emotions were recorded. Employing 16 different combinations of prosody and inventory during the synthesis resulted in 80 synthetic sentences. The results were evaluated by conducting listening tests with 33 naïve listeners. Synthesized anger was recognized with 86.6% accuracy, sadness with 89.1% and happiness with 44.0% accuracy. According to our results, anger was classified as inventory dominant and sadness as contour dominant. Results were not sufficient to make similar conclusions regarding happiness. The highest recognition accuracies were achieved for sentences synthesized by employing pitch contour and diaphones belonging to the same emotion.

Spectral conversion techniques were applied for emotion transformation. Gaussian Mixture Models for a limited set of phonemes, phonemes extracted from the neutral and angry inventories, were built and then the designed transformation algorithms were used to convert neutral phonemes into angry phonemes. First results show that even with a limited training successfully recognized new phonemes can be obtained.

In addition to that, in collaboration with Motion Capture Research at ICT, the multi-modal emotional database was recorded for one male actor and one female actress. The database consisted of spoken emotional sentences and facial expressions.

6. Other Relevant Work Being Conducted and How this Project is Different

A similar copy-synthesis approach has been applied to synthesize emotional speech in Spanish [9]. Recognition rates for Spanish showed that prosodic (supra-segmental) information alone was not enough to portray emotions and that supra-segmental information characterized sadness and surprise, while segmental components became dominant for cold anger and happiness. This conclusion was supported by studies on German emotional speech [10], which stated that 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 increase the recognition results for emotional Austrian German speech [11]. Experiments on synthesizing emotional speech using Japanese emotional corpora with CHATR [12] were also in support of using emotional inventory to synthesize emotional speech.

7. Plan for the Next Year

- Development of conversion algorithms that will enable transformation of one emotional category into the other.
- Recording of new speakers.

- Employment of new synthesis techniques, namely HMM-based speech synthesis techniques [13].

8. Expected Milestones and Deliverables

- Large emotional database
- Algorithms for emotional speech synthesis
- Algorithms for emotional conversion
- A limited domain speech synthesizer capable of manipulating emotions.

9. Member Company Benefits

N/A

10. References

[1] I.R. Murray, J.L. Arnott and E.A Rohwer, "Emotional stress in synthetic speech: progress and future directions”, Speech Communication, 20, 1996, pp. 85-91.

[2] Douglas-Cowie, E., Cowie, R., & Schröder, M. (2000). A new emotion database: Considerations, sources and scope, *ISCA Workshop on Speech and Emotion, Northern Ireland*, p. 39-44.

[3] http://www.festvox.org/

[4] F. Charpentier and E. Moulines. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diaphones. In *Proc. EUROSPEECH*, pages 13-19, 1989.

[5] M. Schröder (2001). Emotional Speech Synthesis - A Review. *Proc. Eurospeech 2001, Aalborg*, Vol. 1, pp. 561-564.

[6] Damasio, A.R., “Descartes’ Error”, Putnam’s Sons, New York, 1994.

[7] Stylianou, Y., Cappe, O., and Moulines, E., “Continuous Probabilistic Transform for Voice Conversion”, *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No.2, March 1998.

[8] Kain, A., and Macon, M.W., “Spectral Voice Conversion for Text-to-Speech Synthesis”, *Proceedings of ICASSP*, vol. 1, pp. 285-288, May 1998.

[9] 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.

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

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

[12] 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.

[13] http://hts.ics.nitech.ac.jp/