Constructing emotional speech synthesizers with limited speech database

Ryosuke Tsuzuki †, Heiga Zen †, Keiichi Tokuda †, Tadashi Kitamura †,
Murtaza Bulut ‡, Shrikanth S. Narayanan ‡

† Department of Computer Science and Engineering, Nagoya Institute of Technology
‡ Speech Analysis & Interpretation Laboratory, University of Southern California
†{ryosuke,zen,tokuda,kitamura}@ics.nitech.ac.jp
‡{mbulut,shri}@sipi.usc.edu

Abstract

This paper describes an emotional speech synthesis system based on HMMs and related modeling techniques. For concatenative speech synthesis, we require all of the concatenation units that will be used to be recorded beforehand and made available at synthesis time. To adopt this approach for synthesizing the wide variety of human emotions possible in speech, implies that this process should be repeated for every targeted emotion making this task challenging and time consuming. In this paper, we propose an emotional speech synthesis technique based on HMMs, especially for the case where only limited amount of training data is available, directly incorporating subjective evaluation results performed on the training data. Listening results performed on the synthesized speech suggest that the proposed technique helps to improve the emotional content of synthesized speech.

1. Introduction

Recently, extensive progress on speech synthesis systems have been reported. To let machines speak like a human, it is important to synthesize speech with various speaking styles and emotions. Unit selection speech synthesis systems, which is one of the most popular speech synthesis systems currently, have shown a significant improvement in the quality of synthesized speech. Several studies for synthesizing emotional speech have already been proposed based on unit selection speech synthesis [1–3]. For the production of emotional speech using concatenative speech synthesis techniques, there are two major methods one might follow. These differ in terms of the kind of the inventory; one may try to induce the targeted emotional effect by modifying the neutral speech units, or try to modify specifically recorded emotional speech samples (or domain specific samples as in [4]) to produce emotional speech. Each of these systems has its own merits and disadvantages. For example in the former approach a great deal of modification at the prosodic level may be necessary; also since the modification of voice quality will not be possible, the required result may not be realizable. Technically, in the later approach voice quality modification will not pose much problem if the coverage of emotional inventory is sufficient; also the potentially less prosodic modifications are required. The real challenge in this approach is however to achieve sufficient coverage in the inventory. Considering the number of the emotions human beings can produce and perceive, and the signal variations that occur due to emotions, it should be obvious that a huge inventory of emotional speech will be necessary. Complex classification and search algorithms and lots of storage place (which is not a problem for personal computers anymore, however, a problem for small embedded devices) will be necessary to select the best available concatenation speech units.

On the other hand, HMM-based speech synthesis systems [5] provide an effective approach for synthesizing speech with various speaking styles and emotions. In this system, voice qualities of synthesized speech can be easily changed by transforming HMM parameters suitably [6–8]. In addition, by using a decision-tree based context clustering technique [9], model parameters with sufficient accuracy can be estimated with limited training data. From this perspective, the advantages offered by the HMM-based approach make it a promising framework for synthesizing speech with emotions.

In an emotional speech synthesis system, natural speech reflecting various (target) emotions is used as training data. However, even with trained voice talents, it is difficult to ensure consistent emotional quality across the entire recording i.e., we expect intra-speaker variability especially if the recording set is large. Furthermore, recording a large amount of speech data is a big burden for the speaker and it takes a very long time. Thus, designing large corpus with desired emotions has many inherent problems. To make emotional synthesized speech from even non-professional speakers, we must construct speech synthesis systems that can synthesize speech even if training data is limited.

This paper proposes a technique for synthesizing emotional speech with limited training data. We first conducted a listening test on the natural speech. Based on obtained evaluation scores in the test, we constructed the emotional contexts of each sentence for context clustering. In this paper, the emotions we chose to work with include anger, happiness, sadness, and neutral. Since the acoustic characteristics of the sentences are evaluated by the listeners, contexts suitable for modeling from a perception-perspective can be obtained. We synthesized two kinds of emotional speech, using contexts based on subjective evaluation and based on emotion presented by a speaker (based on direct elicitation), respectively to compare synthesized speech.

The rest of this paper is organized as follows. Section 2 describes the HMM-based speech synthesis system. Section 3 describes the technique for synthesizing emotional speech. Section 4 describes our experiments, and concluding remarks and the plans for future work are presented in the final section.

2. HMM-based speech synthesis system

2.1. Model

In an HMM-based speech synthesis system, we construct feature vectors using parameters obtained from speech databases.

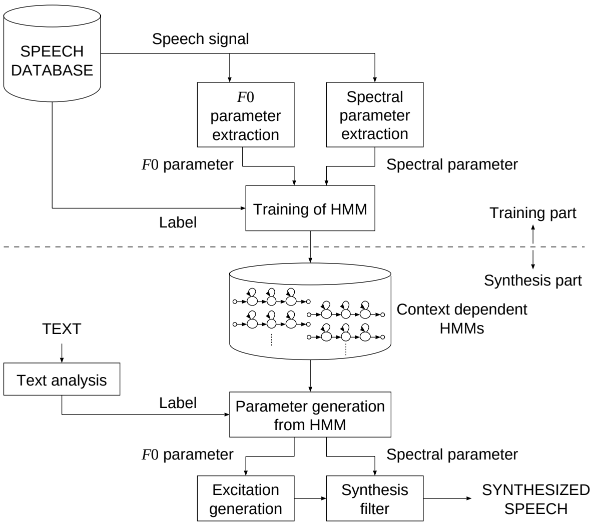

Figure 1: Overview of an HMM-based speech synthesis system.

Table 1: Numbers of sentences.

Neutral	Anger	Happiness	Sadness	Total
74	105	95	89	363

The feature vector consists of mel-cepstral coefficients as spectral parameters, log fundamental frequency (log $F0$) as frequency parameter, and their delta and delta-delta parameters. Sequences of mel-cepstral coefficient vectors are modeled by continuous density HMMs. Log $F0$ sequences are modeled by a Multi-Space probability Distribution HMM (MSD-HMM) [10]. By using MSD-HMM, voiced and unvoiced frames are modeled by continuous and discrete distributions, respectively.

State durations of each HMM are regarded as a multi-dimensional observation and the set of state durations of each HMM is modeled by a multi-dimensional Gaussian distribution.

2.2. System overview

An HMM-based speech synthesis system consists of two parts, training part and synthesis part. The system overview is shown in Fig. 1.

In the training part, first, context-independent HMMs are trained by feature vectors and label boundaries. After converting to context-dependent HMMs, they are estimated by embedded training. However, model parameters with sufficient accuracy cannot be estimated with limited training data. To overcome this problem, we apply a tree-based context clustering technique based on an MDL criterion [9] for the spectrum, $F0$ and state duration. Since each of the feature component (spectrum, $F0$ and duration) has its own contextual factor influence, the distributions for spectral parameter, $F0$ parameter and the state duration are clustered independently.

In the synthesis part, a text to be synthesized is converted to a context-dependent label sequence. Then, a sentence HMM is constructed by concatenating context-dependent HMMs according to the label sequence. State durations of the sentence HMM are determined so as to maximize their probabilities for the state duration densities [11]. According to the obtained state durations, a sequence of mel-cepstral coefficients and $F0$ values including voiced/unvoiced decisions is generated from the sentence HMM by using a speech parameter generation algorithm [12]. Finally, a speech waveform is synthesized directly from the generated parameters by using the MLSA filter [13].

Table 2: Examples of sentences.

Emotional state	Sentence
Neutral	My shoe size is eleven.
Anger	The traffic has been blocked by an accident.
Happiness	My teacher gave me a good grade.
Sadness	I lost the game today.

3. Emotional speech synthesis

3.1. Dataset for Training

For concatenative speech synthesis, we require all of the concatenation units that will be used to be recorded beforehand and made available for processing at synthesis time. Assuming that the emotional set, i.e., the emotions that we are interested in, is decided a priori, the next step is to assure complete phonetic coverage. The need for a complete coverage requires that the context is designed carefully. Specifically, if the emotional recordings will consist of sentences, the sentences should be constructed appropriately; the same is also true if paragraphs will be recorded. Having a specific context to record and the need to obtain specific emotional effects suggests that professional actors and actresses should be employed. In that way, we can be sure that the target emotions have been expressed. To test how the recordings are perceived, listening tests can be performed with native listeners and if the recognition rates are not satisfactory, the recordings can be repeated. Using professional actors in recording the emotional inventory will also minimize the burden of having to subjectively classify emotions after the recording. Another approach for emotional recording, strongly supported by Campbell [14], is recording real life situations with no specifically designed context to record. The advantage of this approach is that more natural emotional expressions will be collected; the disadvantage is that we may need to record a large inventory due to the lack of direct control over the recorded context. In addition to that, the need for the classification of the emotions in the “real life” scene is necessary, an incredibly challenging task.

In this paper, we chose to work with four target emotional states: anger, happiness, sadness, and neutral. First, four different source text scripts (one for each emotional state) were prepared. 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 designed as short declarative sentences. To motivate and focus the speaker, a one or two sentence scenario accompanied each of the source sentences. These brief scenarios were prepared for eliciting happy, sad and angry speech inventories, and were not used for the neutral sentences. The male speaker recorded for this experiment had no professional acting experience. The number of speech data in dataset with each emotional state and examples of emotionally biased sentences are shown in Tab. 1 and Tab. 2, respectively. In such recordings, especially by non-professional speakers, it is difficult to increase the number of sentence any further. This underscores the need for constructing a speech synthesis system which can synthesize speech with various emotional states even when the training data is limited.

3.2. Evaluation of Natural Speech

Recording large amounts of speech with a certain style is very difficult. Although trained voice talent may work well, recording speech with suitable and consistent characteristics of desired emotion has its limitations. Therefore, the corpus with

Table 3: Results of listening test. (Natural Speech)

		Recognition Rate (%)			
		Neu.	Ang.	Hap.	Sad.
Natural Speech	Neu.	68.4	0.0	0.0	31.6
	Ang.	4.8	92.3	2.9	0.1
	Hap.	49.8	0.6	47.1	2.5
	Sad.	20.0	0.0	0.6	79.4

Table 4: Numbers of distributions after context clustering.

Model	Mel-cepstrum	$F0$	Duration
"speaker"	469	1454	1033
"listener"	453	1479	1019

emotional speech may not indeed have the actual emotional styles intended by the "speaker." In modeling the characteristics of emotional speech, it is desirable to use emotional styles based on the perception of the "listener."

Therefore, in this paper, before the speech synthesis experiments, we conducted a subjective listening test for natural speech with 10 listeners who were non-native speakers of English. We expect that it could reduce the effect of contents of each sentence, though there may be cross-cultural effects in the results. Listeners were asked to choose the most suitable emotional state among anger, happiness, sadness, or neutral for all sentences included in training data.

Table 3 shows the results of listening tests for natural speech. We used these results to obtain the contexts about emotions for training as described in 3.3. Since the acoustic characteristics of the sentences are evaluated by the listeners, contexts suitable for modeling from a perception-perspective can be obtained.

However, according to the results of listening tests, the overall recognition rate is 72.4%. Particularly, recognition rate for happiness is lower than 50.0%. These facts show that the difficulty of ensuring consistent emotional quality. To make emotional synthesized speech with limited data spoken by non-professional speaker is very challenging task.

3.3. Contextual Factors

Speech signals are affected by a lot of contextual factors (e.g., phone identify factors, stress-related factors). If more detailed contextual factors are prepared, we can train more accurate acoustic models. In this paper, contextual factors (contexts) based on [5, 15] are taken into account.

In addition, factors of emotion are taken into account as contexts. We prepared two kinds of contexts: one was constructed based on results of the listening tests. First, for each sentence included in training data, we counted the number of votes for each emotion state in the listening test. Using the results, we constructed contexts for each sentence (e.g., if the number of votes for neutral about the sentence was 7, for anger 0, for happiness 1, and for sadness 2, the emotional context for the sentence is "neutral : anger : happiness : sadness = 7:0:1:2."). Another was constructed based on the original intention of the speaker, i.e., emotion state for the sentence presented by a speaker. (e.g., uttered with anger).

4. Experiment

4.1. Synthesis Experiment

Speech signals were sampled at 22.05kHz and windowed by a 24.5-ms Blackman window with a 4.9-ms shift, and then mel-

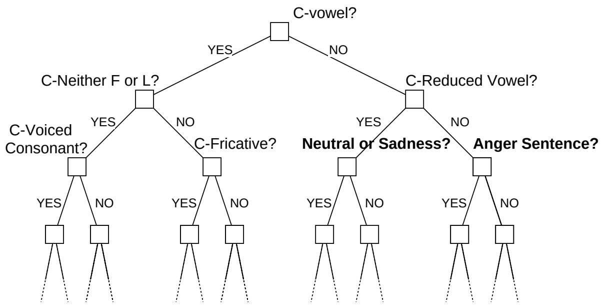

Figure 2: Example of decision tree.("speaker" - $F0$ - 3rd state)

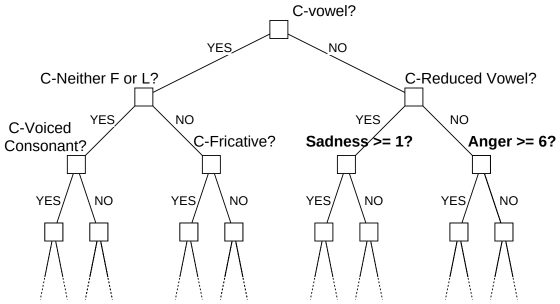

Figure 3: Example of decision tree.("listener" - $F0$ - 3rd state)

cepstral coefficients were obtained by the mel-cepstral analysis. $F0$ parameters were extracted automatically by ESPS get_f0. The resulting $F0$ parameters were not corrected manually. Feature vector consists of spectral and $F0$ parameter vectors. The spectral parameter vector consists of 34 mel-cepstral coefficients including the zeroth coefficient, their delta and delta-delta coefficients. The $F0$ parameter vector consists of log fundamental frequency, its delta and delta-delta. We used 5-state left-to-right HMMs. Each output probability distribution was modeled by a 5-dimensional single Gaussian with diagonal covariance matrix.

In this experiment, all the speech included in the database were used for training HMMs. We constructed models with contexts of speaker's intention ("speaker") and emotion based on subjective evaluation ("listener"), separately. The number of distributions for spectrum models, $F0$ models and state duration models after context clustering are shown in Tab. 4. Examples of decision-trees for "speaker"and "listener"are shown in Fig. 2 and 3, respectively. Context dependent labels except emotional context for training and testing were extracted from texts by Festival Speech Synthesis System [16] without manual correction. In synthesizing speech, we used contexts involving a single emotional state (e.g., "listener" - neutral : anger : happiness : sadness = 0:10:0:0, "speaker" - uttered with anger).

4.2. Evaluation Experiment

One of the biggest challenges in speech technology research tends to be in the evaluation of the results. Being targeted for humans, in most cases the evaluation experiments are performed with humans, which means that the conclusions derived from these results are highly subjective. In addition to that, since it is practically impossible for each researcher to employ the same group of people, technically it is impossible to compare the results in a scientific way. In other words, since each research uses different listener groups, the results are usually biased by the subjective factors directly related to the participants.

In this paper, the forced choice test, where the listeners were required to choose one of the four listed emotional categories (anger, happiness, sadness, and neutral), was used. This test was conducted with 10 listeners. Stimuli presented were generated from “speaker” and “listener” models using contexts involving anger, happiness, sadness, or neutral, respectively. Total number of these categories came to 8. We constructed emotionally neutral sentences not included in the training data for this test (e.g., “I saw somebody jogging under this blazing sun.”). The number of test sentences for each emotion was 20, which was chosen from 50 sentences, randomly for every listener.

Table 5 and 6 show the results of listening test for “speaker” and “listener” models, respectively. Table 7 show correct recognition rates of the results of listening tests. According to these results, the overall recognition rate increases from 53.2% to 58.6% by modeling HMMs based on subjective evaluation (“listener”). Particularly, it is effective for neutral and happiness: recognition rates increased by more than 10%, respectively. In Tab. 3, recognition rates for neutral and happy were lower than those for other emotions. These results show that the proposed technique is more effective in case where it is difficult to recognize desired style in natural speech and we can see that a good level of reproduction of emotional content is achieved by Tab. 7. The modeling power of the HMM-based system is hence promising in cases that require finer details in speech to be captured such as in synthesizing “happiness” which was found to be less successful with just a unit selection scheme [3].

The recognition rate of speech samples synthesized by the proposed technique was lower than that of natural speech (original speech database) calculated from Tab. 3. However, it should be noted that the speech database consists of emotional sentences, while synthesized speech samples consist of neutral sentences. We also conducted a similar experiment with native English listeners. In this case, the recognition rate of natural speech was very high since they might have judged their emotions based on the contents of the sentences. As a result, listener’s judgments were very similar to speaker’s intention, and we could not find a clear difference between “speaker” and “listener.”

Table 5: Results of listening test. (“speaker”)

		Recognition Rate(%)			
		Neu.	Ang.	Hap.	Sad.
Presented Emotion	Neu.	53.0	0.0	0.0	47.0
	Ang.	5.0	81.0	14.0	0.0
	Hap.	73.5	2.5	18.5	5.5
	Sad.	36.5	1.5	1.5	60.5

Table 6: Results of listening test. (“listener”)

		Recognition Rate(%)			
		Neu.	Ang.	Hap.	Sad.
Presented Emotion	Neu.	65.0	1.0	8.0	26.0
	Ang.	2.5	79.5	17.5	0.5
	Hap.	60.5	4.5	29.0	6.0
	Sad.	36.5	0.5	2.0	61.0

Table 7: Results of listening tests. (correct recognition rate(%))

	Neu.	Ang.	Hap.	Sad.	Ave.
Natural Speech (emotional sentences)	68.4	92.3	47.1	79.4	72.4
“speaker”	**53.0**	81.0	**18.5**	60.5	53.2
“listener”	**65.0**	79.5	**29.0**	61.0	58.6

5. Conclusion

In this paper, an HMM-based speech synthesis approach was investigated for emotional speech synthesis. Furthermore, a modeling technique based on subjective evaluation was proposed. Listening test results show that the proposed technique improves the expressiveness of emotions for synthesized speech. Future work includes constructing more effective contexts for emotional speech synthesis. Improvements in the quality of synthesized speech is also a future work.

6. References

[1] A. Iida, N. Campbell, S. Iga, F. Higuchi, and M. Yasumura, “A speech synthesis system for assisting communication,” Proc. of ISCA2000, pp.167–172, 2000.

[2] E. Eide, “Preservation, identification, and use of emotion in a text-to-speech system,” IEEE Speech Synthesis Workshop, Sep. 2002.

[3] M. Bulut, S. S. Narayanan, and A. K. Syrdal, “Expressive speech synthesis using a concatenative synthesizer,” Proc. of ICSLP, pp.1265–1268, Sep. 2002.

[4] L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, “Limited Domain Synthesis of Expressive Military Speech for Animated Characters,” IEEE Speech Synthesis Workshop, Sep. 2002.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” Proc. of EUROSPEECH, vol.5, pp.2347–2350, Sep. 1999.

[6] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, “Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR,” Proc. of ICASSP 2001, vol.2, pp.805–808, May 2001.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Speaker Interpolation in HMM-based speech synthesis system,” Proc. of EUROSPEECH, vol.5, pp.2523–2526, 1997.

[8] K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Eigenvoices for HMM-based speech synthesis,” Proc. of ICSLP, pp.1269–1272, Sep. 2002.

[9] K. Shinoda and T. Watanabe, “Speaker adaptation with autonomous model complexity control by MDL principle,” Proc. of ICASSP, pp.717–720, 1996.

[10] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling,” Proc. of ICASSP, vol.1, pp.229–232, Mar. 1999.

[11] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura “Duration modeling in HMM-based speech synthesis system,” Proc. of ICASSP, vol.2, pp.29–32, 1998.

[12] K. Tokuda, T. Yoshimura, T. Masuko, and T. Kobayashi, “Speech parameter generation algorithms for HMM-based speech synthesis,” Proc. of ICASSP, vol.3, pp.1315–1318, June 2000.

[13] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” Proc. of ICASSP, vol.1, pp.137–140, 1992.

[14] N. Campbell, “Databases of Emotional Speech,” Proc. of ISCA2000, pp.34-38, 2000.

[15] K. Tokuda, H. Zen, and A. W. Black, “An HMM-based speech synthesis system applied to English,” IEEE Speech Synthesis Workshop, Sep. 2002.

[16] *http://www.cstr.ed.ac.uk/projects/festival/*