Session 1pSC

Speech Communication: Synthesis and Recognition of Speech (Poster Session)

Om D. Deshmukh, Cochair
Univ. of Maryland, Electrical and Computer Engineering Dept., A.V. Williams Bldg., College Park, MD 20742

Koichi Shinoda, Cochair
Tokyo Inst. of Technology, Graduate School of Information Science and Engineering, Dept. of Computer Science, 2-12-1 W8-81, Ookayama, Meguro-ku, Tokyo 152-8552, Japan

Contributed Papers

All posters will be on display from 2:00 p.m. to 6:00 p.m. To allow contributors an opportunity to see other posters, contributors of odd-numbered papers will be at their posters from 2:00 p.m. to 4:00 p.m. and contributors of even-numbered papers will be at their posters from 4:00 p.m. to 6:00 p.m.

1pSC1. Improving quality of small body transmitted ordinary speech with statistical voice conversion. Hidehiko Sekimoto, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano (Nara Inst. of Sci. and Technol., 8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192 Japan)

The explosive spread of cellular phones enables us to communicate with each other at any time or place. Although cellular phones are convenient, there are still some problems. For example, it is difficult to send intelligible speech under noisy conditions, which is a fatal problem especially when talking privately using small speech in crowds. To improve the quality of small speech under such situations, a new speech communication style is proposed using a nonaudible murmur (NAM) microphone [Nakajima *et al.*, Eurospeech (2003), pp. IV-2601–2604]. The NAM microphone is robust to eternal noise, although body transmission causes quality degradation. In this paper, to improve the sound quality of small body transmitted ordinary speech (SBTOS), which is small speech recorded with a NAM microphone, two conversion methods that reflect a statistical voice conversion method based on Gaussian mixture models (GMMs) [Toda *et al.* Interspeech (2005), pp. 1957–1960] are conducted. One conversion method is from SBTOS to ordinary speech (SBTPS-to-SP), and the other is from SBTOS to small speech (SBTOS-to-SSP). SBTOS-to-SSP has more consistent correspondence of voiced/unvoiced segments between input and output speech than SBTOS-to-SP. The results of objective and subjective evaluations show that SBTOS-to-SSP outperforms SBTOS-to-SP.

1pSC2. Evaluation of eigenvoice conversion based on Gaussian mixture model. Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano (Grad. School of Information Sci., Nara Inst. of Sci. 8916-5, Takayama-cho, Ikoma-city, Nara, Japan 630-0192)

Eigenvoice conversion (EVC) has been proposed as a new framework of voice conversion (VC) based on the Gaussian mixture model (GMM) [Toda *et al.*, ''Eigenvoice Conversion Based on Gaussian Mixture Model,'' ICSLP, Pittsburgh, Sept. 2006]. This paper evaluates the performance of EVC in conversion from one source speaker's voice to an arbitrary target speakers' voices. This framework trains canonical GMM (EV-GMM) in advance using multiple parallel data sets consisting of utterance pairs of the source and many prestored target speakers. This model is adapted to a specific target speaker by estimating a small number of free parameters using a few utterances of the target speaker. This paper compares spectral distortion between converted and target voices in EVC with conventional VC based on GMM when varying the amount of training data and the number of mixtures. Results show EVC outperforms conventional VC when using small amounts of training data. EVC can effectively train a complex conversion model using the information of many prestored speakers. By contrast, conventional VC needs a large-sized parallel data set for training. It also shows the results of subjective evaluations of speech quality and conversion accuracy for speaker individuality.

1pSC3. Evaluating naturalness of speeches morphed by independently using the interpolation ratios of the time-frequency axes and amplitude. Toru Takahashi, Masanori Morise, and Toshio Irino (Faculty of Systems Eng., Wakayama Univ., 930, Sakaedani, Wakayama, Japan, 640-8510, tall@sys.wakayama-u.ac.jp)

In speech morphing procedure [Kawahara *et al.*, Proc. IEEE-ICASSP 2003, Vol. I, pp. 256–259], two parameters exist for controlling the degree of morphing between the source and target: the interpolation ratios of the time-frequency axes and amplitude. Conventional morphing methods use only a single path in which both ratios are identical. It has, however, been reported that perception of naturalness degrades around a morphing ratio of 0.5. It was assumed that there would be better combination of the ratios with maintaining the quality. It was surveyed that the naturalness contour is within a two-dimensional morphing space. The morphed speech sounds are synthesized in combinations of 11 axis ratios (between 0 and 1 by 0.1 step) and 11 amplitude ratios. Synthetic sounds (121 in all) were presented to subjects to judge the naturalness rate. The optimum path will be described at a presentation to be given at this conference. [This research was partly supported by the ''e-Society Leading Project'' of the Ministry Education, Culture, Sports, Science, and Technology.]

1pSC4. Building an English speech synthetic voice using a voice transformation model from a Japanese male voice. Akemi Iida (School of Media Sci., Tokyo Univ. of Technol., 1404-1, Katakura-cho, Hachiouji, Tokyo, 192-0982, Japan, ake@media.teu.ac.jp), Shimpei Kajima, Kiichi Yasu, Takayuki Arai (Sophia Univ., Tokyo, 102-8554, Japan), and Tsutomu Sugawara (Sophia Univ., Tokyo, 102-8554, Japan)

This work reports development of an English speech synthetic voice using a voice transformation model for a Japanese amyotrophic lateral sclerosis patient as part of a project of developing a bilingual communication aid for this patient. The patient, who had a tracheotomy 3 years ago and had difficulty in speaking, wishes to speak in his own voice in his native language and in English. A Japanese speech synthesis system was developed using ATR CHATR 6 years ago and the authors have worked in developing a diphone-based synthesis using FESTIVAL speech synthesis system and FESTVOX by having the patient read the diphone list. However,

it was not an easy task for the patient to phonate and, moreover, to pronounce words in a foreign language. We therefore used a voice transformation model in FESTIVAL to develop the patient's English speech synthetic voice which enables text-to-speech synthesis. We trained using 30 sentences read by the patient and those synthesized with an existing FESTIVAL diphone voice created from a recording of a native English speaker. An evaluation including a listening experiment was conducted and the result of this voice conversion showed that the synthesized voice successfully reflected the voice quality of the patient.

1pSC5. An MRI-based time-domain speech synthesis system. Tatsuya Kitamura, Hironori Takemoto, Parham Mokhtari (NICT/ATR Cognit. Information Sci. Labs., 2-2-2 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0288, Japan), and Toshio Hirai (Arcadia, Inc., Minoushi, Osaka, 562-0003, Japan)

A speech synthesis system was developed based on Maedas method [S. Maeda, Speech Commun. **1**, 199–229 (1982)], which simulates acoustic wave propagation in the vocal tract in the time domain. This system has a GUI interface that allows fine control of synthesis parameters and timing. In addition, the piriform fossae were included to the vocal tract model, resulting in antiresonances in speech spectra at the frequency region from 4 to 5 kHz. The system can produce all the Japanese phonemes using vocal tract area-functions (VTAFs) extracted from 3-D cine-MRI obtained during production of VCV or CVCV sequences for a male speaker. The system can be used to synthesize Japanese sentences with high naturalness and intelligibility by concatenating segmental units and controlling the glottal source using the GUI interface. Since a time-varying VTAF is obtained by interpolating between VTAFs, the dataset size of the system is significantly smaller than that of corpus-based speech synthesizers. The speaker-specific VTAFs and inclusion of the piriform fossae permit us to reproduce speaker-specific spectral shapes, not only the lower formants but also higher frequency regions that contribute to the perception of speaker individualities. [Work supported by NICT, SCOPE-R, and Grant-in-And for Scientific Research of Japan.]

1pSC6. Database size and naturalness in concatenative speech synthesis. H. Timothy Bunnell, James T. Mantell, and James B. Polikoff (Speech Res. Lab., A. I. duPont Hospital for Children, 1600 Rockland Rd., Wilmington, DE 19803)

Unit concatenation TTS systems seek to maximize perceived naturalness by minimizing the amount of signal processing applied to the recorded speech on which they are based. To generate distinct suprasegmentals for a given segmental sequence (e.g., to convey variation in focus or emotion), it is necessary to record and store multiple instances of the same segments that vary in fundamental frequency and voice quality. At the expense of naturalness, concatenative systems can store a minimal segmental inventory and synthesize suprasegmental factors by manipulating $f0$ and voice quality via signal processing. Classic diphone synthesis (where only a single instance of each diphone sequence is stored) represents the limiting case of this strategy. The present study explores aspects of the trade-off between perceived naturalness and segmental inventory size using the ModelTalker TTS system. Twenty-five speakers each recorded about 1650 utterances. From these, databases were constructed that limited the maximum number of alternate diphones usable for synthesis in five conditions to 1, 5, 10, 20, and 40. Sentences were synthesized from these databases using either full or no $f0$ control. Results of listening tests wherein subjects rate the naturalness of each sentence will be presented. [Work supported by NIDCD.]

1pSC7. Optimization of target cost weights in concatenative speech synthesis with very short segments of 5-ms duration. Toshio Hirai (Arcadia, Inc., 3-1-15 Nishishoji, Mino, Osaka 5620003, Japan, thirai@arcadia.co.jp)

If a concatenative speech synthesis system uses more short speech segments, it increases the potential to generate natural speech because the concatenation variation becomes greater. Recently, a synthesis approach was proposed in which very short (5 ms) segments are used [T. Hirai and S. Tenpaku, 5th ISCA Speech Synthesis Workshop, pp. 37–42 (2004), http://www.ssw5.org/]. In that approach, the target cost (how close a database segment is to desired segment) was defined as the simple average of the root mean squares (rms's) of the difference between the features of the database segment and the desired segment for simplicity. Therefore, it has been expected to optimize the weight of each rms. A Japanese speech database was used to optimize the weights and to evaluate its effects. For corpus construction, 150 utterances were selected from the database. Ten other utterances were selected randomly for generation of feature time series as natural targets for synthesis. Half of them were used for optimization using many weight combinations in synthesis to determine the optimal weight set that shows the minimum concatenation distortion. Distortion for the other half of the utterance synthesis with optimized weights was reduced 34.4% compared to the former approach.

1pSC8. Improving Japanese syllable unit selection using search space with small acoustical features variance. Takaaki Moriyama and Seiichi Tenpaku (Arcadia, Inc., 3-1-15, Nishishoji, Mino, Osaka, 5620003 Japan, taka@arcadia.co.jp)

Usually in concatenative text-to-speech systems, suitable units are selected according to acoustical features. However, a method to reduce calculation costs has been proposed, which is using syllable notations and positions along with the mora length [Murakami *et al.*, Trans. IEICE **J85-D-II**(7), 1157–1165 (2002)]. An enlargement of the corpus to synthesize longer phrases lowers the quality of synthesized speech because of increased feature variance within syllable units with identical notation. Methods to produce detailed groups in which the feature's variance is smaller than syllable units, and to synthesize speech using these groups, are proposed. To produce the groups, a mel-cepstrum distance is used as a measure for evaluation. In this method, the quality of synthesized speech that is given an ideal group sequence might improve, as the distance between units is close within the same group. As a first step to evaluate the proposed method, five groups were generated from a Japanese corpus; then, speech that was synthesized to replace one unit in natural speech to others in the same groups was examined. Results of this listening experiment indicated that an average of the naturalness for these speech types was greater than 80%.

1pSC9. Toward hidden Markov model-based spontaneous speech synthesis. Tatsuya Akagawa, Koji Iwano, and Sadaoki Furui (Dept. of Comput. Sci., Tokyo Inst. of Technolology, 2-12-1-W8-77 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan, akagawa@furui.cs.titech.ac.jp)

This paper investigates several issues of spontaneous speech synthesis. Although state-of-the-art synthesis systems can achieve highly intelligible speech, their naturalness is still low. Therefore, much work must still be done to achieve the goal of synthesizing natural, spontaneous speech. To model spontaneous speech using a limited amount of data, we used an HMM-based speech synthesizer based on three features: cepstral features modeled by HMMs, and duration and fundamental frequency features modeled using Quantification Theory Type I. The models were trained with approximately 17 min of spontaneous lecture speech, from a single speaker, which was extracted from the Corpus of Spontaneous Japanese (CSJ). For comparison, utterances by the same speaker, reading a transcription of the same lecture, were used to train analogous models for read speech. Spontaneity of the synthesized speech was evaluated by subjective pair comparison tests. Results obtained from 18 subjects showed that the preference score for the synthesized spontaneous speech was significantly