Session 1aSCb

Speech Communication: Speech Processing Potpourri (Poster Session)

Jeffrey Berry, Cochair
jjberry@email.arizona.edu

Contributed Papers

All posters will be on display from 9:20 a.m. to 12:20 p.m. To allow contributors an opportunity to see other posters, contributors of odd-numbered papers will be at their posters from 9:20 a.m. to 11:00 a.m. and contributors of even-numbered papers will be at their posters from 11:00 a.m. to 12:20 p.m.

1aSCb1. Entropy coding for training deep belief networks with imbalanced and unlabeled data. Jeffrey Berry (University of Arizona, Department of Linguistics, Tucson, AZ 85721, jjberry@email.arizona.edu), Ian Fasel (University of Arizona, School of Information: Science, Technology and Arts, Tucson, AZ 85721), Luciano Fadiga (Italian Institue of Technology, Department of Robotics, Brain and Cognitive Sciences, Genoa, Italy 16163), and Diana Archangeli (University of Arizona, Department of Linguistics, Tucson, AZ 85721)

Training deep belief networks (DBNs) is normally done with large data sets. In this work, the goal is to predict *traces* of the surface of the tongue in ultrasound images of the mouth during speech. Performance on this task can be dramatically enhanced by pre-training a DBN jointly on human-supplied traces and ultrasound images, then training a modified version of the network to predict traces from ultrasound only. However, hand-tracing the entire dataset of ultrasound images is extremely labor intensive. Moreover, the dataset is highly imbalanced since many images are extremely similar. This work presents a bootstrapping method which takes advantage of this imbalance, iteratively selecting a small subset of images to be hand-traced, then (re)training the DBN, making use of an entropy-based diversity measure for the initial selection. With this approach, a three-fold reduction in human time required to trace an entire dataset with human-level accuracy was achieved.

1aSCb2. Voice search optimization using weighted finite-state transducers. Yuhong Guo, Ta Li, Yujing Si, Jielin Pan, and Yonghong Yan (Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, guoyuhong@hccl.ioa.ac.cn)

Voice search system can provide users with information according to their spoken queries. However, as the most important module in this system, the high word error rate of the automatic speech recognition (ASR) part degrades the whole system's performance. Moreover, the runtime efficiency of the ASR also becomes the bottleneck in the large scale application of voice search. In this paper, an optimized weighted finite-state transducer (WFST) based voice search system is proposed. A weighed parallel silence short-pause model is introduced to reduce both the final transducer size and the word error rate. The WFST network is optimized as well. The experimental results show that, the recognition speed of proposed system outperforms the other recognition system at the equal word error rate and the miracle error rate is also significantly reduced. This work is partially supported by the National Natural Science Foundation of China (No's. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

1aSCb3. Hybrid low delay frame loss concealment in an MDCT based audio codec. Zhibin Lin (Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, Jiangsu, China, zblin@nju.edu.cn), Ming Wu (State Key Laboratory of Acoustics, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190,China), Jing Lu, and Xiaojun Qiu (Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, Jiangsu, China)

By combining tonal-dominant and noise-dominant signal frame loss concealment (FLC) approaches, a hybrid low delay FLC method is proposed for an modified discrete cosine transform (MDCT) based codec. Based on the observations that the phase of the MDCT-MDST (modified discrete sine transform) coefficients of tonal-dominant signals decreases linearly with the increase of the frame index and the amplitude keeps unchanged, the tonal-dominant signal FLC approach uses the frame interpolation to estimate the phase and magnitude of the MDCT-MDST coefficients of the lost frame while the noise-dominant signal FLC method implements a modified shaped-noise insertion. Both objective and subjective test results show that the proposed technique provides better performance than the existing methods for music signals and voiced speech signals.

1aSCb4. Multi-band speech recognition using band-dependent confidence measures of blind source separation. Atsushi Ando, Hiromasa Ohashi (Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan, atsushi.ando@g.sp.m.is.nagoya-u.ac.jp), Sunao Hara (Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan), Norihide Kitaoka, and Kazuya Takeda (Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan)

One of the main applications of Blind Source Separation (BSS) is to improve performance of Automatic Speech Recognition (ASR) systems. However, conventional BSS algorithm has been applied only to speech signals as a pre-processing approach. In this paper, a closely coupled framework between FDICA-based BSS algorithm and speech recognition system is proposed. In the source separation step, a confidence score of the separation accuracy for each frequency bin is first estimated. Subsequently, by employing multi-band speech recognition system, acoustic likelihood is calculated from the estimated BSS confidence scores and Mel-scale filter bank energy. Therefore, our proposed method can reduce ASR errors which caused by separation errors in BSS and permutation errors in ICA, as in the conventional approach. Experimental results showed that our proposed method improved word accuracy of ASR by approximately 10%.