Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 243–264
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Maximum likelihood sub-band adaptation for robust speech recognition

Donglai Zhu [a,b,*], Satoshi Nakamura [a], Kuldip K. Paliwal [a,c], Renhua Wang [b]

[a] *ATR Spoken Language Translation Research Labs, Japan*
[b] *University of Science and Technology of China, China*
[c] *School of Microelectronic Engineering, Griffith University, Australia*

Received 5 January 2004; received in revised form 15 February 2005; accepted 15 February 2005

Abstract

Noise-robust speech recognition has become an important area of research in recent years. In current speech recognition systems, the Mel-frequency cepstrum coefficients (MFCCs) are used as recognition features. When the speech signal is corrupted by narrow-band noise, the entire MFCC feature vector gets corrupted and it is not possible to exploit the frequency-selective property of the noise signal to make the recognition system robust. Recently, a number of sub-band speech recognition approaches have been proposed in the literature, where the full-band power spectrum is divided into several sub-bands and then the sub-bands are combined depending on their reliability. In conventional sub-band approaches the reliability can only be set experimentally or estimated during training procedures, which may not match the observed data and often causes degradation of performance. We propose a novel sub-band approach, where frequency sub-bands are multiplied with weighting factors and then combined and converted to cepstra, which have proven to be more robust than both full-band and conventional sub-band cepstra in our experiments. Furthermore, the weighting factors can be estimated by using maximum likelihood adaptation approaches in order to minimize the mismatch between trained models and observed features. We evaluated our methods on AURORA2 and Resource Management tasks and obtained consistent performance improvement on both tasks.

Keywords: Robust speech recognition; Sub-band; Adaptation

1. Introduction

It is well known that current ASR systems don't work as well as humans. Existing recognizers are extremely sensitive to channel variability and

* Corresponding author. Address: Dept. Computer Science, The University of Hong Kong, Pokfulam road, Hong Kong.
E-mail addresses: dlzhu@cs.hku.hk, donglai.zhu@ustc.edu (D. Zhu), satoshi.nakamura@atr.co.jp (S. Nakamura), k.paliwal@me.gu.edu.au (K.K. Paliwal), rhw@ustc.edu.cn (R. Wang).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.006

additive background noise and require careful preprocessing. However, humans are able to achieve excellent recognition accuracy in these cases. Fletcher and his colleagues (Fletcher, 1953; Allen, 1994) suggested that in human auditory perception, the linguistic message gets decoded independently in different frequency sub-bands and the final decoding decision is based on merging the decisions from the sub-bands. Some other experiments studied human performance on filtered (low-pass, high-pass, band-pass and band-reject filtered) speech, in order to gain a better understanding of human speech perception or to find the contribution of different parts of the speech spectrum for perception. Miller and Niely (1955) showed that humans can achieve high recognition rates for narrow-band speech signals. Kryter (1960) showed that humans can combine the perception coming from narrow sub-bands to enhance intelligibility. Riener et al. (1992) and Warren et al. (1995) showed that high intelligibility can be maintained for band-pass filtered speech signals. They also concluded that humans possess processing mechanisms that are able to employ limited spectral regions that can enhance comprehension under difficult listening conditions. Lippmann (1996) showed that humans can recognize speech signals produced by severe band-reject filtering. The results discussed above show that humans can recognize speech signals with limited spectral cues and can easily integrate acoustic cues from different frequency regions for speech perception.

Noise-robust speech recognition has become an important area of research in recent years. In current speech recognition systems, the Mel-frequency cepstrum coefficients (MFCCs) are used as recognition features. When the speech signal is corrupted by narrow-band noise, the entire MFCC feature vector gets corrupted and it is not possible to exploit the frequency-selective property of the noise signal to make the recognition system robust. Recently, a number of sub-band speech recognition approaches have been proposed in the literature, where the full-band power spectrum is divided into several sub-bands and then the sub-bands are combined depending on their reliability. In conventional sub-band approaches the reliability can only be set experimentally or estimated during training procedures, which may not match the observed data and often causes degradation of performance.

Two modes of sub-band approaches have been proposed: parallel sub-band (PSB) and concatenating sub-band (CSB) (Hermansky et al., 1996; Bourlard and Dupont, 1996, 1997; Tibrewala and Hermansky, 1997; Cerisara et al., 1998, 2000; Okawa et al., 1998; Paliwal and Chen, 2000). In both modes, features are extracted from the sub-band spectra independently. If the cepstrum is used as the feature, the extraction procedure is as follows: firstly, the frequency spectrum of the speech signal is divided into sub-bands; secondly, for each sub-band, the spectrum is converted to a cepstrum. There are several free parameters in the procedure, e.g., the number of sub-bands and the frequency boundaries of the sub-bands. They can be adjusted to appropriate values for given tasks (Tibrewala and Hermansky, 1997). For the PSB, the sub-band features are modeled independently, and the likelihood scores of the sub-bands are combined at some segmental levels (phonemes, words and sentences, etc.). Results showed that the PSB can improve the recognition performance for speech signals corrupted by band-limited noises (Bourlard and Dupont, 1997), but may exhibit poorer performance when the additive background noise has wide bands (Tibrewala and Hermansky, 1997). An issue for the PSB is the sub-band recombination. Mainly, three methods have been studied for it. The first is the weighted average method (Bourlard and Dupont, 1996, 1997; Cerisara et al., 1998; Okawa et al., 1998). It produces the overall probability based on an arithmetic or geometric average of the sub-band probabilities. The contribution of each sub-band is weighted by its local signal-to-noise ratio (SNR), or by its reliability. The second is the neural network approach (Hermansky et al., 1996; Tibrewala and Hermansky, 1997). Multi-layer perceptions (MLPs) may be trained to merge the sub-band probabilities to estimate the probability of all possible combinations of the sub-bands. The third is the full combination. The probabilities of different combinations of different sub-bands are combined using a linear method, with each weight