Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 707–715

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency

Hui Yin [a,b,*], Volker Hohmann [a,c], Climent Nadeu [a]

[a] *TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain*
[b] *Department of Electronic Engineering, Beijing Institute of Technology, Beijing 100081, China*
[c] *Medizinische Physik, Universität Oldenburg, Germany*

Received 16 July 2009; received in revised form 8 April 2010; accepted 17 April 2010
Available online 7 May 2010

Abstract

Most of the features used by modern automatic speech recognition systems, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) coefficients, represent spectral envelope of the speech signal only. Nevertheless, phase or frequency modulation as represented in recent perceptual models of the peripheral auditory system might also contribute to speech decoding. Furthermore, such features can be complementary to the envelope features. This paper proposes a variety of features based on a linear auditory filterbank, the Gammatone filterbank. Envelope features are derived from the envelope of the subband filter outputs. Phase/frequency modulation is represented by the subband instantaneous frequency (IF) and is used explicitly by concatenating envelope-based and IF-based features or is used implicitly by IF-based frequency reassignment. Speech recognition experiments using a standard HMM-based recognizer under both clean training and multi-condition training are conducted on a Chinese mandarin digits corpus. The experimental results show that the proposed envelope and phase based features can improve recognition rates in clean and noisy conditions compared to the reference MFCC-based recognizer.

Keywords: Gammatone filterbank; Instantaneous frequency; Speech recognition

1. Introduction

The human auditory system is a highly complex sensory system. Studying its structure and function is of great importance for a better understanding of sensory processes in general and of hearing pathologies and their treatment in particular. Furthermore, these studies may also provide conceptual ideas for the design of information-processing systems that mimic human capabilities (Munkong and Juang, 2008). It is well-known, e.g., that automatic speech recognition (ASR) systems perform far less reliably than a human listener under adverse conditions such as those encountered in a noisy environment or over changing transmission channels. Thus, integrating human auditory processing characteristics into ASR systems might improve the recognition rate significantly. Aim of this study is to investigate the potential of spectro-temporal signal representations inspired by auditory processing for improving ASR.

Most of the features used by modern speech recognition systems, such as mel-frequency cepstral coefficients (MFCC) and perceptual linear predictive (PLP) coefficients, just represent spectro-temporal envelope of the signal. In realistic environments, however, feature sets that represent only a limited subset of the properties of the signal may easily be corrupted. Using the information obtained from other signal properties and combining them

[*] Corresponding author at: Department of Electronic Engineering, Beijing Institute of Technology, Beijing 100081, China. Tel.: +86 13810514236; fax: +86 01068915218.

E-mail addresses: hchhuihui@gmail.com, hchhuihui@126.com (H. Yin), volker.hohmann@uni-oldenburg.de (V. Hohmann), climent.nadeu@upc.edu (C. Nadeu).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.04.008

was found to be important to minimize the effects of the environment on ASR (Kubo et al., 2008). In particular, Kubo et al. (2008) found that frequency modulation information can be complementary to the envelope features, which might reflect the specific sensitivity of the human auditory system to frequency modulated sounds like police and ambulance siren. Recent models of auditory processing represent spectral envelope and amplitude modulation as well as several features that are related to the phase spectrum, e.g., frequency modulation. As a step towards including these features into ASR feature sets, envelope and phase based representations derived from the output of an auditory filterbank will be investigated in this study. In particular, the subband instantaneous frequency (IF) is used as a phase based feature. Although the representations employed here are not directly derived from a complete auditory model, they represent aspects of the information possibly used by the auditory system. Phase based features will therefore be regarded as auditory features for the purpose of this study.

The application of auditory models to speech recognition has already been studied extensively. Kleinschmidt et al. (2001) used the model of auditory perception (PEMO) as a front end for ASR. Schluter et al. (2007) introduced the linear auditory Gammatone filterbank for large-vocabulary speech recognition. Holmberg et al. (2007) assessed a detailed model of auditory peripheral processing by means of ASR. Recent work on auditory models shows that some of the nonlinear effects of active cochlear processing might be simulated using the IF estimated in auditory frequency subbands as a parameter for gain control (Hohmann and Kollmeier, 2006). Subband IF has already been applied to speech analysis and ASR in recent work. Stark and Paliwal (2008) introduced the IF deviation spectrum which exhibits both pitch and formant structure similar to the magnitude spectrum. Potamianos and Maragos (2001) discussed the use of general time–frequency distributions as features for ASR in the context of hidden Markov classifiers. Short-time averages of quadratic operators, e.g., energy spectrum, generalized first spectral moments and short-time averages of the instantaneous frequency are compared to the standard front-end features and applied to ASR. The average instantaneous frequency (AIF) and average log envelope (ALE) has been used for ASR (Kumaresan et al., 2003; Wang et al., 2003). Kubo et al. (2008) combined a pseudo-instantaneous frequency analyzer with an amplitude modulation analyzer. Alsteris and Paliwal (2007) made the effort of incorporating information from the short-time phase spectrum into a feature set for ASR, and investigated two intermediate representations: the group delay function (GDF) and the instantaneous frequency distribution (IFD). Haque et al. (2009) applied a zero-crossing auditory model as a pre-processing front end to ASR in noisy environments. Dimitriadis et al. (2005) measured the amount of amplitude and frequency modulation that exists in speech resonances and integrated it into the acoustic features for ASR. These features are mainly the first-order statistics (mean values) of the demodulated instantaneous signal, and combined with MFCCs to achieve lower error rate.

All the literature above proves that features derived from the phase spectrum may be supplementary to the magnitude spectrum, although in many of the reported experiments the proposed feature sets performed worse than the MFCCs, or just achieved a modest improvement. Larger improvements were obtained when the new features were combined with standard MFCCs. The features based on instantaneous frequency were proven to be useful to ASR. Nevertheless, such features are in most papers limited to first-order statistics of the IF, i.e., mean of IF values, and the combination of the IF-based features and MFCCs is just a concatenation on feature level or weighted combination on probability level.

Several reasons for the modest success of features derived from auditory models and the phase spectrum in ASR applications might be hypothesized. Auditory-model and phase based representations employ generally a very high spectral and temporal resolution. As a consequence, feature patterns represent both speech-cues relevant to ASR and speech-cues irrelevant to ASR (e.g., speaker-specific cues, mood, stress, etc.) in a mixed form. A possible performance gain from the relevant features could be countervailed by distracting effects of irrelevant features (undesired variability generated by the irrelevant features). Linear discriminant analysis (LDA) or similar methods might alleviate partly the problem of relevant and irrelevant information being mixed in the auditory features, but a generally applicable method to separate both types of information is yet to be investigated. Furthermore, a high temporal resolution comes with a high variability of the data, and requires a high dimensionality of the feature sets. To alleviate these problems, temporal averaging of features across time frames was used in most studies as mentioned above. Although this leads to features that can be handled more easily by HMM recognizers, the potential benefits of auditory features might partly get lost. Further studies of higher-order statistics of IF and other combination methods are therefore suitable to explore the potential of auditory features in more detail. In particular, variance and entropy of the IF calculated in time frames are proposed as novel features in this study. It is hypothesized that these measures of intra-frame variation add useful information and still can be combined with HMM decoders.

In this study, a linear auditory Gammatone filterbank is used to do front-end speech processing. Basic features based on energy/envelope and instantaneous frequencies (IF) are derived from the outputs of the Gammatone filterbank. The IF information is integrated into basic features in various ways. In addition to the mean IF, variance and entropy of the IF are investigated as novel features. Furthermore, IF-based frequency reassignment as proposed in (Plante et al., 1998; Potamianos and Maragos, 1996; Gardner and Magnasco, 2005) is investigated. In this