Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 736–752
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A hierarchical framework for spectro-temporal feature extraction

Martin Heckmann [a,*], Xavier Domont [a,b], Frank Joublin [a], Christian Goerick [a]

[a] Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany
[b] Technische Universität Darmstadt, Control Theory and Robotics Lab, D-64283 Darmstadt, Germany

Available online 11 August 2010

Abstract

In this paper we present a hierarchical framework for the extraction of spectro-temporal acoustic features. The design of the features targets higher robustness in dynamic environments. Motivated by the large gap between human and machine performance in such conditions we take inspirations from the organization of the mammalian auditory cortex in the design of our features. This includes the joint processing of spectral and temporal information, the organization in hierarchical layers, competition between coequal features, the use of high-dimensional sparse feature spaces, and the learning of the underlying receptive fields in a data-driven manner. Due to these properties we termed the features as hierarchical spectro-temporal (HIST) features. For the learning of the features at the first layer we use Independent Component Analysis (ICA). At the second layer of our feature hierarchy we apply Non-Negative Sparse Coding (NNSC) to obtain features spanning a larger frequency and time region. We investigate the contribution of the different subparts of this feature extraction process to the overall performance. This includes an analysis of the benefits of the hierarchical processing, the comparison of different feature extraction methods on the first layer, the evaluation of the feature competition, and the investigation of the influence of different receptive field sizes on the second layer. Additionally, we compare our features to MFCC and RASTA-PLP features in a continuous digit recognition task in noise. On a wideband dataset we constructed ourselves based on the Aurora-2 task, as well as on the actual Aurora-2 database. We show that a combination of the proposed HIST features and RASTA-PLP features yields significant improvements and that the proposed features carry complementary information to RASTA-PLP and MFCC features.

Keywords: Spectro-temporal; Auditory; Robust speech recognition; Image processing; Learning; Competition; Hierarchical

1. Introduction

Humans have the astonishing ability to preserve stable representations even in a dynamic environment. In contrast to automatic speech recognition systems additional background noise and changes in the transmission channel have only minor effects on human performance (Lippmann, 1997; Sroka and Braida, 2005). Hence better understanding the underlying processes in humans bears the potential to yield better recognition systems.

Unfortunately, our knowledge of the auditory processing in the mammalian brain is still quite limited. The visual system is for example already much better understood (King and Nelken, 2009). This is also expressed in the wealth of corresponding computational models of the visual system. On the other hand, there are several studies highlighting important similarities between the two systems. Sur et al. (1988) showed that newborn ferrets whose retinal nerves were rerouted to the auditory part of the thalamus, sometimes called the gateway to the cortex (Crick, 1984), where later able to respond to visual stimuli via their auditory cortex. This at least demonstrates a high plasticity of these areas during development if not a strong similarity in functional organization. Despite significant differences between auditory and visual processing in the brain common, modality independent processing principles are usually assumed (Read et al., 2002; King and Nelken, 2009). Different authors have shown that at least at the level of the receptive fields in the primary visual and

* Corresponding author. Tel.: +49 69 8901 1755.
E-mail addresses: martin.heckmann@honda-ri.de (M. Heckmann), xavier.domont@rtr.tu-darmstadt.de (X. Domont), frank.joublin@honda-ri.de (F. Joublin), christian.goerick@honda-ri.de (C. Goerick).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.006

auditory cortices these similarities can be found (Schreiner and Calhoun, 1994; de Charms et al., 1998; Shamma, 2001). Measurements in the primary auditory cortex of different animals revealed its spectro-temporal organization, i.e. the receptive fields are selective to modulations in the time-frequency domain. The corresponding receptive fields have, as in the visual cortex, Gabor-like shapes. The above mentioned findings suggest that modeling principles known from image processing can beneficially be transferred to auditory tasks.

Traditionally, speech features mainly took inspirations from psychoacoustic findings and thereby relied in most cases on independent spectral (Hermansky, 1990; Hermansky and Morgan, 1994; Flynn and Jones, 2008; Haque et al., 2009) or temporal representations (Hermansky and Sharma, 1998).

In recent years also features inspired by above mentioned similarities between visual and auditory processing, i.e. features capable of directly capturing spectro-temporal variations, were developed. In his seminal work Kleinschmidt (2002) introduced the usage of 2 D Gabor features for speech recognition in noise. This was followed by others, also employing Gabor features on similar tasks, including speech vs. non-speech discrimination (Mesgarani et al., 2006; Meyer and Kollmeier, 2008; Sherry and Zhao, 2008). In (Elhilali and Shamma, 2006) a spectro-temporal respresentation based on Gabor filters was used for source separation. Ezzat et al. (2007) randomly selected spectro-temporal patches of the target word from the training set and then used these as features for keyword spotting.

The framework for the extraction of spectro-temporal speech features we present here takes many inspirations from the visual object recognition system of Wersing and Körner (2003) and is an extension of our previous work (Domont et al., 2007, 2008). In contrast to the previously mentioned approaches and other models in the literature we integrated additional processing principles which are also inspired by the mammalian sensory cortex. The processing in the sensory cortices seems to be organized in a hierarchical fashion. This has been stated for the visual (Hubel and Wiesel, 1965; Felleman and Van Essen, 1991) and auditory cortex (Rauschecker, 1998; Read et al., 2002; Scott et al., 2003). Based on this principle we propose a hierarchical framework consisting of two layers.[1] Features in the first layer extract local information. In the second layer the results of these local features are combined to form more complex features. The hierarchical processing and the construction of more complex and at the same time more specific features leads to a substantial increase in the number of features. A trend also observed in the human brain where approximately 3500 inner hair cells are present in the cochlea and about 100,000,000 neurons in the auditory cortex (Dusan and Rabiner, 2005).

In general, when dealing with spectro-temporal features developing methods for the selection of the relevant features is a key issue. Kleinschmidt (2002) already proposed to use a so-called "feature finding neural network" to select the features yielding the best recognition rates. The unsupervised learning of sparse spectro-temporal representations was proposed in (Klein et al., 2003; Behnke, 2003). Cho and Choi (2005) investigated how such learned representations can be applied to the task of sound classification. We follow this idea in that we learn the receptive fields on both layers of our hierarchy with unsupervised learning rules. Another important property of the mammalian sensory cortices is the competition between coequal features. To model this we implemented a Winner-Take-Most competition between the features on the first layer.

The following sections will describe our framework in more detail and will evaluate its performance in comparison to conventional speech features. In Section 2 we give an overview on our framework. Section 3 describes the preprocessing we apply to the speech signal prior to the feature extraction. The learning of the receptive fields and the extraction of the spectro-temporal features is detailed in Section 4. Section 5 presents recognition results we obtain on a noisy digits task. Based on this task this section also evaluates the contribution of the different elements of our framework to the overall performance. Finally, in Section 6 we discuss the results we obtain and possible improvements.

2. Overview

The key elements of our hierarchical feature extraction framework are depicted in Fig. 1. The first step performs a preprocessing of the speech signal and mainly consists of a transformation into the frequency domain and an enhancement of the formant structure. Based on this we calculate local spectro-temporal features. This step is followed by a competition between these local features. Together these two steps constitute the first layer of our framework. On the second layer the local features are combined to form complex features, spanning larger time and frequency regions. The final steps are an orthogonalization of the features via a Principal Component Analysis (PCA) and recognition of the feature stream with a Hidden Markov Model (HMM) based recognizer. Strictly speaking we do not consider the last two steps as being part of our framework. However, they are necessary to evaluate the feature extraction. Due to their hierarchical organization we termed the features resulting from the proposed framework as hierarchical spectro-temporal (HIST) features.

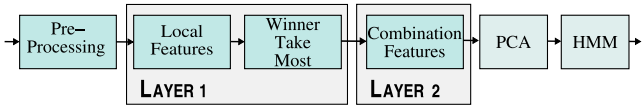

Fig. 1. Overview of the feature extraction process.

[1] Behnke (2003) already suggested a hierarchical speech feature extraction framework but did not evaluate the resulting features in respect to what information they extract and how this could be used for speech processing.