Speech Communication 108 (2019) 15–32

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition

Dimitri Palaz[a,b,c,1,*], Mathew Magimai-Doss[b], Ronan Collobert[d,b,1]

[a] Speech Graphics Ltd., Edinburgh, United Kingdom
[b] Idiap Research Institute, Martigny, Switzerland
[c] Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[d] Facebook A.I. Research, Menlo Park, CA, USA

ARTICLE INFO

Keywords:
Automatic speech recognition
Hidden Markov models
Deep learning
Feature learning
Artificial neural networks
Convolution neural networks
Hybrid HMM/ANN

ABSTRACT

In hidden Markov model (HMM) based automatic speech recognition (ASR) system, modeling the statistical relationship between the acoustic speech signal and the HMM states that represent linguistically motivated subword units such as phonemes is a crucial step. This is typically achieved by first extracting acoustic features from the speech signal based on prior knowledge such as, speech perception or/and speech production knowledge, and, then training a classifier such as artificial neural networks (ANN), Gaussian mixture model that estimates the emission probabilities of the HMM states. This paper investigates an end-to-end acoustic modeling approach using convolutional neural networks (CNNs), where the CNN takes as input raw speech signal and estimates the HMM states class conditional probabilities at the output. Alternately, as opposed to a divide and conquer strategy (i.e., separating feature extraction and statistical modeling steps), in the proposed acoustic modeling approach the relevant features and the classifier are jointly learned from the raw speech signal. Through ASR studies and analyses on multiple languages and multiple tasks, we show that: (a) the proposed approach yields consistently a better system with fewer parameters when compared to the conventional approach of cepstral feature extraction followed by ANN training, (b) unlike conventional method of speech processing, in the proposed approach the relevant feature representations are learned by first processing the input raw speech at the sub-segmental level ($\approx$ 2 ms). Specifically, through an analysis we show that the filters in the first convolution layer automatically learn "in-parts" formant-like information present in the sub-segmental speech, and (c) the intermediate feature representations obtained by subsequent filtering of the first convolution layer output are more discriminative compared to standard cepstral features and could be transferred across languages and domains.

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems typically divide the task of recognizing speech into several sub-tasks, which are optimized in an independent manner (Rabiner and Juang, 1993; Bourlard and Morgan, 1994). Specifically, as a first step, acoustic feature observations, such as Mel frequency cepstral coefficients (MFCCs) or perceptual linear prediction cepstral features (PLPs), are extracted from the short-term speech signal based on speech production and speech perception knowledge. Next, likelihood of subword units, which are typically based on phonemes, are estimated using a statistical model that captures the relationship between the features and the subword units in either generative or discriminative manner. Finally, given the likelihood estimates of the subword units, the best matching word hypothesis is searched by integrating lexical and syntactical constraints.

Recent advances in machine learning have shown that systems can be trained in an end-to-end manner, i.e. systems where every step is *learned* simultaneously, taking into account all the other steps and the final task of the whole system. It is typically referred to as *deep learning* (Hinton et al., 2006; Bengio et al., 2007), mainly because such architectures are usually composed of many layers (supposed to provide an increasing level of abstraction), compared to classical "shallow" systems. As opposed to "divide and conquer" approaches presented previously where each step is independently optimized, deep learning approaches are often claimed to lead to more optimal systems. As they alleviate the need of finding the right features by instead

[*] Corresponding author at: Speech Graphics Ltd., Edinburgh, United Kingdom.
E-mail addresses: dpalaz@speech-graphics.com (D. Palaz), mathew@idiap.ch (M. Magimai-Doss), ronan@collobert.com (R. Collobert).
[1] All research was conducted at the Idiap Research Institute, before Dimitri Palaz joined Speech Graphics and Ronan Collobert joined Facebook AI Research.

https://doi.org/10.1016/j.specom.2019.01.004
Received 29 June 2016; Received in revised form 10 January 2019; Accepted 28 January 2019
Available online 30 January 2019

training a stack of features in an end-to-end manner, for a given task of interest.

While there is a good success record of such approaches in the computer vision (LeCun et al., 1998; Krizhevsky et al., 2012; He et al., 2015) or text processing fields (Collobert et al., 2011b), deep learning approaches for speech recognition has largely focused on the classifier step, where a neural network with many hidden layers is typically trained to classify subword units (Hinton et al., 2012). These systems still rely on standard short-term spectral-based feature extraction. The training optionally can involve pre-training schemes. In such a case, it is referred to as deep belief neural networks (DBNs) otherwise deep neural networks (DNNs).

More recently, there has been efforts toward modeling raw speech signal with little or no pre-processing (Jaitly and Hinton, 2011; Palaz et al., 2013b; Tüske et al., 2014; Golik et al., 2015; Sainath et al., 2015). Towards that, as one of the first efforts, we proposed a novel approach based on convolution neural networks (Palaz et al., 2013b). In this approach, the input to the CNN is raw speech signal. The neural network architecture consists of two stages: a feature learning stage consisting of several convolution layers followed by a classifier stage consisting of multilayer perceptron, which are jointly learned by minimizing a cost function based on relative entropy. Phoneme recognition studies on the TIMIT corpus showed that the proposed approach is capable of achieving performance comparable to or better than the standard approach of extraction of cepstral features followed by ANN training. Subsequent works in the ASR community have explored different architectures. For instance, in Tüske et al. (2014) use of DNNs was investigated. It was found that such an acoustic model yields inferior system when compared to standard acoustic modeling. In a subsequent follow up work (Golik et al., 2015), it was found that addition of convolution layers at the input helps in improving the system performance and reducing the performance gap w.r.t standard acoustic modeling technique. In Sainath et al. (2015), a composite architecture referred to as CLDNN was investigated, where the raw speech signal is fed as input to CNNs, the CNN stage output is subsequently processed by a bidirectional long-short term memory (BLSTM) stage and fed into a DNN stage to classify phones. All these stages are jointly learned. This approach was found to yield performance comparable to the case where the input to CLDNN is log filter bank energies.

An aspect that differentiates our approach from the subsequent works (Tüske et al., 2014; Golik et al., 2015; Sainath et al., 2015) is the manner in which the input speech signal is processed by the ANN. More precisely, in Sainath et al. (2015) the first CNN layer consisted of 40 filters following the standard practise in MFCC or PLP cepstral feature extraction for 8 kHz bandwidth speech signal; the filter lengths were set to 25 ms (400 samples) following standard short-term processing practise; and were initialized with Gammatone impulse response, i.e. based on auditory knowledge. In Tüske et al. (2014) the input to DNN was non-overlapping 10 ms speech signal. They also investigated initialization of the first layer of the DNN with Gammatone impulse response. In Golik et al. (2015), the input convolution layer consisted of 128 filters and the filter lengths were set to 16 ms (256 ms). In our approach, however, the filter length and the number of filters in the first convolution layer is not set a priori, rather they are determined during the training process through cross validation. In other words, the ANN learns how the speech signal should be blocked or windowed as short segments and spectrally processed for phone classification. As a consequence of this flexibility, as we will see later in the present paper, the processing of speech at the input of ANN in the proposed approach considerably departs from the current understanding of short-term speech processing.

The present paper builds on our previous works (Palaz et al., 2013b; 2015b; 2015a) along two directions,

1. From phoneme recognition to automatic speech recognition: a first set of fundamental question that arises is: does the findings on phoneme recognition task scale well across speech recognition task across different languages and domains? In that respect, the contributions of the present paper are: (1) we first benchmark the proposed approach on TIMIT corpus by extending our previous study (Palaz et al., 2013b) to the standardized protocol of classifying 183 phones output (61 phones × 3 states) and using phone bigram for decoding; (2) We then present investigations on large vocabulary continuous speech recognition task on a variety of corpora that differ in terms of languages. Specifically, we extend our previous investigations on WSJ English (Palaz et al., 2015b) to Swiss French and Swiss German on Mediaparl corpus that contains spontaneous speech. Our studies show that the architecture of three convolution layers followed by a multilayer layer perceptron originally developed in the context of phoneme sequence recognition task scales well for continuous speech recognition tasks and consistently yields a better system than conventional cepstral feature-based system for all the investigated corpora.
2. Understanding the learned features: as it would be seen in the ASR studies the proposed approach yields a system that performs better than the system based on conventional approach with considerably less parameters. Thus, a second set of questions that arise are: what information is the neural network learning and how it is learning? Since the features are learned along with the classifier automatically from the data, yet another question that arises is: are these features domain or language dependent? To understand these aspects, we first analyze the first convolution layer. We present a novel signal theoretic approach to understand the information that is collectively modeled by the first convolution layer. This analysis shows that: (i) the proposed approach transforms the speech signal at sub-segmental level (about 2 ms) as opposed to conventional approach of transforming the signal at segmental level (20–30 ms), (ii) unlike auditory motivated filter banks, the learned set of filters are not of constant Q nature, and (iii) as opposed to an adhoc approach presented in our earlier work (Palaz et al., 2015a), through the novel signal theoretic interpretation, we show that the first convolution layer learns a spectral dictionary that models *in-parts* formant-like information in the envelop of magnitude spectrum of sub-segmental speech. We then focus the analysis on the classifier stage, where we show that the learned features are more discriminative than the conventional cepstral features and can be classified well with a simple classifier such as a single layer perceptron. Finally, through cross-domain and cross-lingual studies we show that the learned features could be transferred across languages and domains.

The remainder of the paper is organized as follows. Section 2 presents a background on hybrid HMM/ANN ASR, feature extraction and use of deep neural networks, and motivates the present work. Section 3 presents the architecture of the CNN-based system. Section 4 presents the recognition studies and Section 5 presents the analyses. Section 6 presents a discussion and concludes the paper.

2. Background

This section briefly introduces standard hybrid HMM/ANN ASR system. It then presents a concise survey on two aspects of acoustic modeling: features and ANN-based classifier upon which the present paper focuses on.

2.1. Hybrid HMM/ANN ASR system

As presented in Fig. 1, hybrid HMM/ANN based ASR system is composed of three parts: features extraction, classification and decoding. In the first step, input features $\mathbf{x}_t$ at time t are extracted from the short-term signal $\mathbf{s}_t$. They are then given as input to an artificial neural network (ANN). In literature, ANNs with different architectures have been proposed such as multilayer perceptron (MLP) (Bourlard and Morgan, 1994), time delay neural networks (Waibel et al., 1989) which

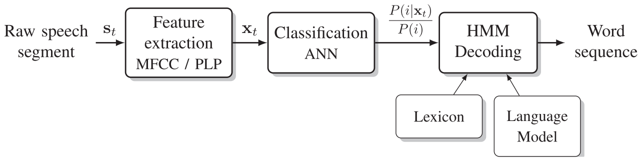


Fig. 1. Hybrid HMM/ANN system. $\mathbf{s}_t$ denotes the input speech segment, $\mathbf{x}_t$ here denotes cepstral features and i denotes a phoneme class.

is also referred to as convolutional neural networks, recurrent neural networks (RNN) (Robinson, 1994; Graves et al., 2013). The ANN estimates the class conditional probabilities $P(i|\mathbf{x}_t)$ for each phone class $i \in \{1, \dots, I\}$. The emission probabilities $p_e(\mathbf{x}_t|i)$ of the HMM states are scaled likelihoods which, as given below, are obtained by dividing the ANN output by the prior probability of the class $P(i)$,

$$p_e(\mathbf{x}_t|i) \propto \frac{p(\mathbf{x}_t|i)}{p(\mathbf{x}_t)} = \frac{P(i|\mathbf{x}_t)}{P(i)} \; \forall i \in 1, \dots, I. \tag{1}$$

The prior class probability $P(i)$ is often estimated by counting on the training set. The phone classes $\{1, \dots, I\}$ can be either context-independent phones or clustered context-dependent HMM states, typically obtained by decision tree based state clustering and tying. Depending upon that the system is referred to as either context-independent phone-based ASR system or context-dependent phone-based ASR system, respectively.

Given the scaled-likelihood estimates, a phonetic lexicon and a language model, the decoder finally infers the best matching word hypothesis through search.

2.2. Feature and classifier

Speech signal is a non-stationary signal. Alternately, the statistical characteristics of the signal change over time due to various reasons such as the speech sound being produced, speaker variation, emotional state variation. In the case of ASR, we are primarily interested in the characteristic of the speech signal that relates to or differentiates the speech sounds. In other words, the primary goal is to estimate statistical evidence about speech sounds given the speech signal. To achieve that, guided by statistical pattern recognition techniques, originally the problem has been split into two steps, namely, feature extraction and modeling of the features by a statistical classifier.

Speech coding studies in telephony have shown that speech can be processed as short segments, transformed, transmitted and reconstructed while keeping the intelligibility or message intact (Rabiner and Schafer, 1978). In particular, the studies have shown that short-term speech signal could be considered as output of a linear time invariant vocal tract filter excited by periodic or aperiodic vibration of vocal cords (Rabiner and Schafer, 1978). Furthermore, speech intelligibility can be preserved by preserving the envelop structure of the short-term spectrum of speech signal, which characterizes the vocal tract system (Schroeder and Atal, 1985). The two most common spectral-based features Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) and perceptual linear prediction (PLP) cepstral coefficients (Hermansky, 1990) are built on those aspects while integrating the knowledge about speech and sound perception.

As illustrated in Fig. 2(a), the extraction of MFCC or PLP feature involves: (1) transformation of short-term speech signal to frequency domain; (2) filtering the spectrum based on critical bands analysis, which is derived from speech perception knowledge; (3) applying a non-linear operation; and (4) applying a transformation to get reduced dimension decorrelated features. This process only models the local spectral level information on a short time window typically of 20–30 ms. The information about speech sound is spread over time. To model the temporal information intrinsic in the speech signal dynamic features are computed by taking approximate first and second derivative of the static features (Furui, 1986).

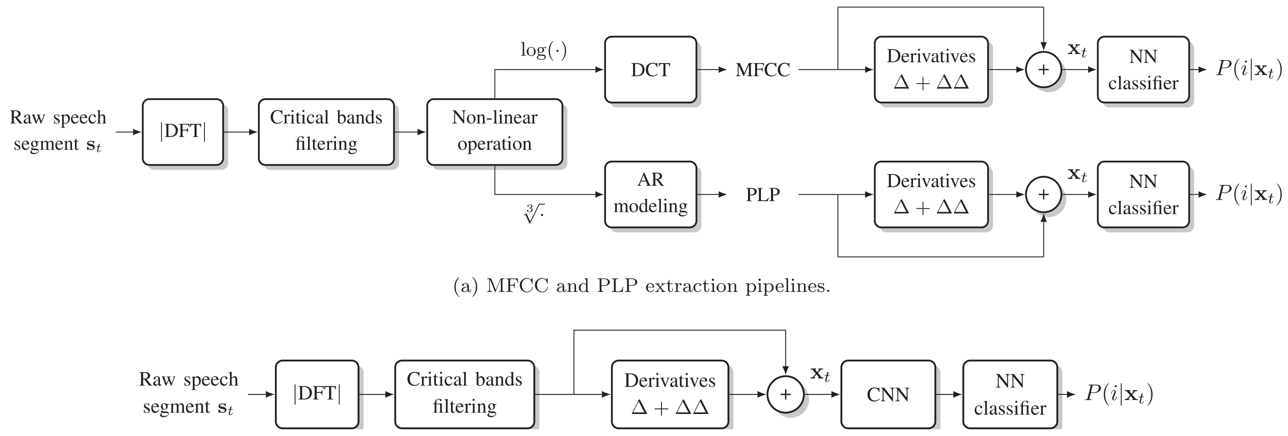
To estimate statistical evidence of speech sounds given the speech signal, the cepstral features are modeled by classifiers such as k-means (or vector quantization), Gaussian mixture models, ANNs, k-nearest neighbor. In the beginning of the hybrid HMM/ANN theory, the ANNs typically had single hidden layer. There were two particular reasons for that. First, it has been shown theoretically that ANN with single hidden layer is an universal approximator (Hornik et al., 1989). Second, both acoustic and computing resources were then limited. In recent years, with the advancements in computing and availability of increased amount of acoustic resources, it has been shown that ANNs with deep architecture, i.e. with multiple hidden layers, can yield better systems (Hinton et al., 2006; Seide et al., 2011; Dahl et al., 2012; Hinton et al., 2012).

2.3. Motivation

The standard acoustic modeling mechanism can be seen as a process of applying transformations guided by prior knowledge about speech production and perception on the speech signal, and subsequent modeling of the resulting features by a statistical classifier. More recently, inspired by the success of deep learning approaches in the fields of text processing and vision (Collobert et al., 2011b; Krizhevsky et al., 2012; He et al., 2015) towards building end-to-end systems as well as by the success of DNNs in ASR, researchers have started questioning the intermediate step of feature extraction. In that direction, several studies have been carried where filter bank or critical band energies estimated from the short-term signal instead of cepstral features are used as input of convolutional neural networks based systems (Abdel-Hamid et al., 2012; Sainath et al., 2013; Swietojanski et al., 2014) or short-term magnitude spectrum is used as input to the DNN (Mohamed et al., 2012; Lee et al., 2009). Fig. 2(b) illustrates a case where, instead of transforming the critical band energies into cepstral features, the critical band energies and its derivatives are fed as input to the ANN.

In this article, as opposed to the idea of applying spectral transform and then learning feature and classifier, we go one step further where the neural network also learns short-term windowing and spectral processing along with the features and the classifier for phone classification. More precisely, in this approach the raw speech signal is input to an ANN that classifies speech sounds. During training the neural network learns the appropriate window size and filtering process that operates on the signal to model the relevant features and the classifier for phone classification. The output of the trained neural network is then used as emission probabilities of HMM states as done in hybrid HMM/ANN approach. Such an approach can not only be motivated by recent advances in machine learning (Collobert et al., 2011b; Krizhevsky et al., 2012) but also from previous works in the speech literature, which have investigated methods to directly model raw speech signal for speech recognition, as presented below.

The first initiative towards directly modeling the raw speech signal was inspired by speech production model, i.e. an observed speech signal can be seen as an output of a time varying filter excited by a time varying source. Specifically, one of the first theoretical work in that direction by Poritz (1982) was inspired by linear prediction techniques, which can deconvolve the excitation source and the vocal tract system through time domain processing. Poritz's work was later revisited as switching autoregressive HMM (Ephraim and Roberts, 2005), and more recently in the framework of switching linear dynamical systems (Mesot and Barber, 2008). These techniques were investigated in an isolated word recognition setup where word-based models are trained. It was found that in comparison to HMM-based ASR system using cepstral features these approaches yield performance comparable under clean conditions and significantly better performance under noisy conditions (Mesot and Barber, 2008). In Sheikhzadeh and Deng (1994), an approach to model raw speech signal was proposed using auto-regressive HMM. In this ap-



(a) MFCC and PLP extraction pipelines.

(b) Typical CNN based pipeline using Mel filter bank [25, 26]

Fig. 2. Illustration of several feature extraction pipelines. |DFT| denotes the magnitude of the discrete Fourier transform, DCT denotes the magnitude of the discrete cosine transform, AR modeling stands for auto-regressive modeling, Δ and $\Delta\Delta$ denote the first and second order derivatives across time, respectively. $P(i|\mathbf{x}_t)$ denotes the conditional probabilities for each input frame $\mathbf{x}_t$, for each label i. It is worth noting that typically, in addition to $\mathbf{x}_t$, the input to the ANN also consists of features from preceding and following frames.

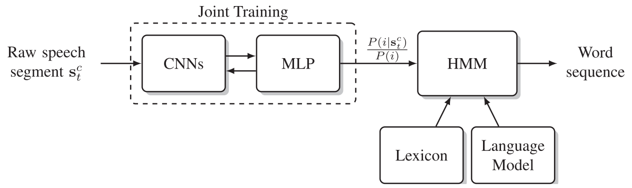


Fig. 3. Overview of the proposed CNN-based approach.

proach, each sample of the speech signal is an observation, as opposed to a vector of speech samples in the approach proposed in Poritz (1982). Each state models the observed speech sample as a linear combination of past samples plus a ”driving sequence” (assumed to be a Gaussian *i.i.d* process). The potential of the approach was demonstrated on classification of speaker-dependent discrete utterances consisting of 18 highly confusable stop consonant-vowel syllables. However, their gain compared to conventional cepstral-based features is not clear, and they were never studied on continuous speech recognition task.

More recently, use of raw speech signal as input to discriminative systems has been investigated. In that direction, combination of raw speech and cepstral features in the framework of support vector machine has been investigated for noisy phoneme classification (Yousafzai et al., 2009). Feature learning from raw speech using neural networks-based systems has been investigated in Jaitly and Hinton (2011). In this approach, the learned features are post-processed by adding their temporal derivatives and used as input for another neural network. Thus, this approach still follows the “divide and conquer” approach. In comparison to these approaches, as presented in the following section, in our approach the features and the classifier are learned in an end-to-end manner to estimate the phone class conditional probability $P(i|\mathbf{x}_t)$ in Eq. (1).

3. Proposed CNN-based approach

We present a novel acoustic modeling approach based on convolutional neural networks (CNN), where the input speech signal $\mathbf{s}_t^c = \{s_{t-c} \dots s_t \dots s_{t+c}\}$ is a segment of the raw speech signal taken in context of 2c frames spanning w_{in} milliseconds. The input signal is processed by several convolution layers and the resulting intermediate representations are classified to estimate $P(i|\mathbf{s}_t^c)$, $\forall i$, as illustrated in Fig. 3. $P(i|\mathbf{s}_t^c)$ is subsequently used to estimate emission scaled-likelihood $p_e(\mathbf{s}_t^c|i)$. As presented in Fig. 4, the network architecture is composed of several filter stages, followed by a classification stage. A filter stage involves a convolutional layer, followed by a temporal pooling layer and a non-linearity, *HardTanh*(·). The number of filter stages is determined during training. The feature stage and the classifier stage are jointly trained using the backpropagation algorithm.

The proposed approach employs the following understandings:

1. Speech is a non-stationary signal. Thus, it needs to be processed in a short-term manner. Traditionally, in the literature guided by Fourier spectral theory and speech analysis-synthesis studies the short-term window size is set as 20–40 ms. The proposed approach follows the general idea of short-term processing. However, the size of the short-term window is a hyper-parameter which is determined during training.
2. Feature extraction is a filtering operation. This can be simply observed from the fact that generic operations such as Fourier transform, discrete cosine transform etc. are filtering operations. In conventional speech processing, the filtering takes place in both frequency (e.g. filter-bank operation) and time (e.g. temporal derivative estimation). The convolution layers in the proposed approach build on these understandings. However, aspects such as the number of filtering layers and their parameters are determined and learned during training, respectively.
3. Though the speech signal is processed in a short-term manner, the information about the speech sounds is spread across time. In conventional approach, the information spread across time is modeled by estimating temporal derivatives and by using contextual information, i.e. by appending features from preceding and following frames, at the classifier input. In the proposed approach the intermediate representations feeding into the classifier stage are estimated using long time span of input speech signal, which is again determined during training. Alternately, w_{in} is a hyper-parameter.

In essence the proposed approach with minimal assumptions or prior knowledge learns to process the speech signal to estimate $P(i|\mathbf{s}_t^c)$.

3.1. Convolutional layer

While “classical” linear layers in standard MLPs accept a fixed-size input vector, a convolution layer is assumed to be fed with a sequence of T vectors/frames: $\{\mathbf{y}_1 \dots \mathbf{y}_t \dots \mathbf{y}_T\}$. As illustrated in Fig. 5, a convolutional layer applies the same linear transformation over each successive (or interspaced by dW frames) windows of kW frames. In this work, $\mathbf{y}_t$ is either a segment of input raw speech $\mathbf{s}_t^c$ (for the first convolution layer) or an intermediate representation output by the previous convolution

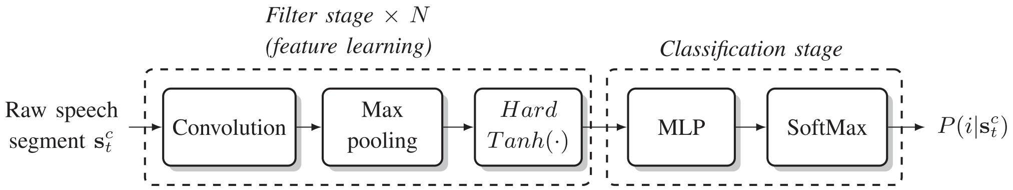


Fig. 4. Overview of the convolutional neural network architecture. Several stages of convolution/pooling/HardTanh might be considered. Our network included three stages. The classification stage can have multiple hidden layers.

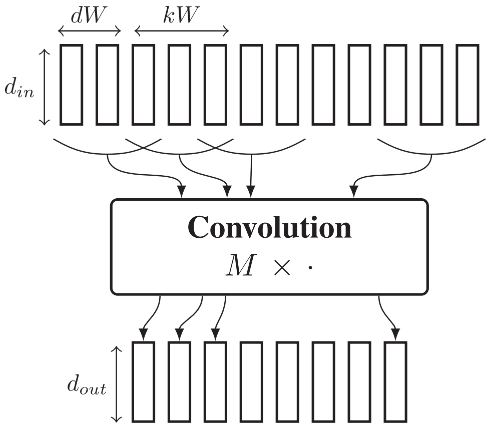


Fig. 5. Illustration of a convolutional layer. d_{in} and d_{out} are the dimensions of the input and output frames. kW is the kernel width (here $kW = 3$) and dW is the shift between two linear applications (here, $dW = 2$).

layer. Formally, the transformation at frame t is written as:

$$M\begin{pmatrix} \mathbf{y}_{t-(kW-1)/2} \\ \vdots \\ \mathbf{y}_{t+(kW-1)/2}, \end{pmatrix}, \tag{2}$$

where M is a $d_{out} \times (kW \cdot d_{in})$ matrix of parameters, d_{in} denotes the dimension of each input frame and d_{out} denotes the output dimension of each frame. In other words, d_{out} filters (rows of the matrix M) are applied to the input sequence.

3.2. Max-pooling layer

This kind of layers perform local temporal max operations over an input sequence. More formally, the transformation at frame t is written as:

$$\max_{t-(kW_{mp}-1)/2 \le k \le t+(kW_{mp}-1)/2} \mathbf{y}_k^d \qquad \forall d, \tag{3}$$

with $\mathbf{y}$ being the input and $d \in \{1, \cdots d_{out}\}$. These layers increase the robustness of the network to minor temporal distortions in the input. They also bring some level of invariance to the phase of the signal, as a phase difference between two signals can be seen as a temporal shift.

3.2.1. Non-linearity

This kind of layer applies a non-linearity to the input. In this work, we use the *HardTanh* layer, defined as:

$$HardTanh(x) = \begin{cases} -1 & \text{if } x < -1 \\ x & \text{if } -1 \le x \le 1 \\ 1 & \text{if } x > 1 \end{cases} \tag{4}$$

This layer is a hard version of the hyperbolic tangent. It has the advantage of being cheaper to compute while keeping the generalization performance of the exact tangent (Collobert, 2004). It is worth mentioning that other types of non-linearities such as, rectified linear unit (ReLU) (Nair and Hinton, 2010; Zeiler et al., 2013) can also be applied (e.g., see (Palaz, 2016, Chapter 6)).

3.3. Softmax layer

The *Softmax* (Bridle, 1990) layer interprets network output scores $f_i(\mathbf{s}_t^c)$ as conditional probabilities, for each class label i:

$$P(i|\mathbf{s}_t^c) = \frac{e^{f_i(\mathbf{s}_t^c)}}{\sum_j e^{f_j(\mathbf{s}_t^c)}}. \tag{5}$$

3.4. Network training

The network parameters θ are learned by maximizing the log-likelihood $\mathcal{L}$, given by:

$$\mathcal{L}(\theta) = \sum_t \log(P(i_t|\mathbf{s}_t^c, \theta)), \tag{6}$$

for each speech segment $\mathbf{s}_t^c$ and its corresponding label i_t, over the whole training set, with respect to the parameters of each layer of the network. Defining the `logadd` operation as:

$$\operatorname*{logadd}_j(z_j) = \log\left(\sum_j e^{z_j}\right). \tag{7}$$

The log-likelihood $\mathcal{L}_t$ of frame t can be expressed as:

$$\mathcal{L}_t = \log(P(i_t|\mathbf{s}_t^c)) = f_{i_t}(\mathbf{s}_t^c) - \operatorname*{logadd}_j(f_j(\mathbf{s}_t^c)), \tag{8}$$

where $f_{i_t}(\mathbf{s}_t^c)$ described the network score for the frame label i_t. Maximizing this likelihood is performed using the stochastic gradient ascent algorithm (Bottou, 1991).

3.5. Illustration of a trained network

In the proposed approach, in addition to the number of hidden units in each hidden layer of the classification stage, the filter stage has number of hyper-parameters, namely, time span of input speech signal w_{in} used to estimate $P(i|\mathbf{s}_t^c)$, number of convolution layers, kernel or temporal window width kW at input of each convolution layer, dW shift of the temporal window at the input of each convolution layer, max pooling kernel width kW_{mp} and shift of max pooling kernel dW_{mp}. In the present work, all of these hyper-parameters are determined during training based on frame level classification accuracy on validation data.

Fig. 7 illustrates the trained feature stage of the proposed CNN approach on the TIMIT corpus. The details of the training can be found in the following Section 4. The filter stage has three convolution layers and it takes a window of 250 ms speech signal w_{in} as input to estimate $P(i|\mathbf{s}_t^c)$ every 10 ms. The figure also illustrates the temporal information κ modeled by the output of each convolution layer and the temporal shift δ. Briefly, the first convolution layer models in a fine grain manner the changes in the signal characteristics over time, i.e. processes 1.8 ms of speech ($kW = 30$ samples) every 0.6ms ($dW = 10$ samples). The subsequent convolution layers then filter and temporally integrate the output of the first convolution layer to yield an intermediate feature representation that is input to the classifier stage, which eventually yields an estimate of $P(i|\mathbf{s}_t^c)$.

It is worth pointing out that the dimensionality of the intermediate representation at the feature learning stage output depends upon the number of convolution stages and the max-pooling kernel width. As it can be seen that max-pooling is done without temporal overlap. So, at each convolution stage, in addition to filtering minor temporal distortions, max-pooling operation acts as a down sampler.

4. Recognition studies

In this section, we present automatic speech recognition studies to show the potential of the proposed approach. We compare it against the conventional approach of spectral-based feature extraction followed by ANN training on different tasks and languages, namely, (a) TIMIT phoneme recognition task, (b) Swiss French Mediaparl task and (c) Swiss German Mediaparl task. The Wall Street Journal (WSJ) 5k task (Palaz et al., 2015b) is also reported for the sake of completeness. The objective of these studies is to demonstrate the potential of the proposed end-to-end acoustic modeling approach by comparing it against the standard cepstral feature-based acoustic modeling for estimating phoneme class posterior probability.

The reminder of the section is organized as follows. Section 4.1 presents the different datasets and setup used for the studies. Section 4.2 presents the different systems that are trained and evaluated. Section 4.3 presents the recognition studies.

4.1. Databases and setup

4.1.1. TIMIT

The TIMIT acoustic-phonetic corpus (Garofolo et al., 1993) consists of 3696 training utterances (sampled at 16 kHz) from 462 speakers, excluding the SA sentences. The validation set consists of 400 utterances from 50 speakers. The core test set is used to report the results. It contains 192 utterances from 24 speakers, excluding the validation set. Experiments are performed using 61 phoneme labels, with three states, for a total of 183 targets as in Mohamed et al. (2009). After decoding, the 61 hand labeled phonetic symbols are mapped to 39 phonemes, as presented in Lee and Hon (1989).

4.1.2. Wall street journal

The Wall Street Journal (WSJ) corpus is an English corpus consisting of read microphone speech (Paul and Baker, 1992). The SI-284 set of the corpus is formed by combining data from WSJ0 and WSJ1 databases (Woodland et al., 1994). The set contains 36,416 sequences sampled at 16 kHz, representing around 80 h of speech. 10% of the set is taken as the validation set. The Nov'92 set is selected as test set. It contains 330 sequences from 10 speakers. The dictionary is based on the CMU phoneme set, 40 context-independent phonemes. We obtain 2776 clustered context-dependent (cCD) units, i.e. tied-states, by training a context-dependent HMM/GMM system with decision tree-based state tying using HTK (Young et al., 2002). We use the bigram language model provided with the corpus. The test vocabulary contains 5000 words.

4.1.3. Mediaparl

MediaParl is a bilingual corpus (Imseng et al., 2012) containing data (debates) in both Swiss German and Swiss French which were recorded at the Valais parliament in Switzerland. Valais is a state which has both French and German speakers with high variability in local accents specially among German speakers. Therefore, MediaParl provides a real-speech corpus that is suitable for ASR studies. In our experiments, audio recordings with 16 kHz sampling rate are used.

The Swiss German part of the database, referred to as *MP-DE*, is partitioned into 5955 sequences from 73 speakers for training (14 h), 876 sequences from 8 speakers for validation (2 h) and 1692 sequences from 7 speakers (4 h) for test. 1101 tied-states are used in the experiments, following the best system available on this corpus (Razavi et al., 2014). The vocabulary size is 16,755 words. The dictionary is provided in the SAMPA format with a phone set of size 57 (including sil) and contains all the words in the training, validation and test set. A bigram language model is used.

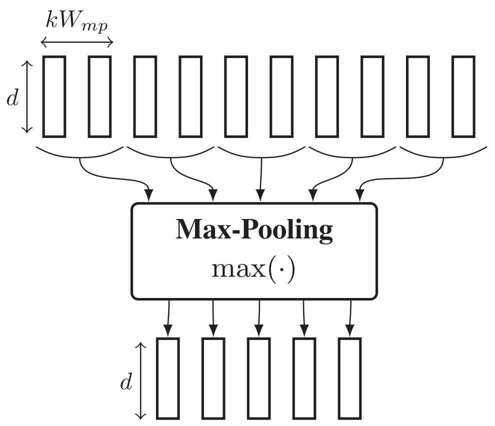


Fig. 6. Illustration of max-pooling layer. *kW* is the number of frames taken for each max operation (here, $kW_{mp} = 2$ and $dW_mp = 2$) and d represents the dimension of input/output frames (which are equal).

The Swiss French part of the database, referred to as *MP-FR*, is partitioned into 5471 sequences from 107 speakers for training (14 h), 646 sequences from 9 speakers for validation (2 h) and and 925 sequences from 7 speakers (4 h) for test. 1084 tied-states are used in the experiments, as presented in Razavi and Magimai.-Doss (2014). The vocabulary size is 12,035 words. The dictionary is provided in the SAMPA format with a phone set of size 38 (including sil) and contains all the words in the training, validation and test sets. A bigram language model is used.

4.2. Systems

In this section, for each task studied, we present the details of the conventional spectral feature based baseline systems (Section 4.2.1) and the proposed CNN-based system using raw speech signal as input (Section 4.2.2). All neural networks were initialized randomly and trained using the Torch7 toolbox (Collobert et al., 2011a). The HTK toolbox (Young et al., 2002) was used for the HMMs and the cepstral features extraction.

4.2.1. Conventional cepstral feature based system

On each task, we have two baseline hybrid HMM/ANN systems which differ in terms of ANN architecture. More precisely, 1 hidden layer MLP (denoted as ANN-1H) based system and 3 hidden layers MLP (denoted as ANN-3H) based system. These ANNs estimate $P(i|\mathbf{x}_t)$, where $\mathbf{x}_t$ is a cepstral feature vector at time frame t. The details of the baseline systems for the different tasks are as follows,

- TIMIT: We treat the one hidden layer MLP based system and the three hidden layers MLP based system without pre-training i.e. random initialization reported in Mohamed et al. (2012), Fig. 6 as the baseline systems. Our motivation in doing so is that they are one of the best cepstral feature-based systems without use of adaptation methods reported in the literature on this task. In these systems, the inputs to the MLPs were 39 dimensional MFCC features ($c_0 - c_{12} + \Delta + \Delta\Delta$) with five frames preceding and five frames following context (i.e. input dimension 39 × 11). ANN-1H has 2048 nodes in the hidden layer and ANN-3H has 1024 nodes in each of the three hidden layers.
- WSJ: We trained an ANN-1H and an ANN-3H to classify 2776 tied-states. The inputs to the MLP are 39 dimensional MFCC features ($c_0 - c_{12} + \Delta + \Delta\Delta$) with four frames preceding and four frames following context (i.e. input dimension 39 × 9). The MFCC features are computed with a frame size of 25 ms and a frame shift of 10 ms.

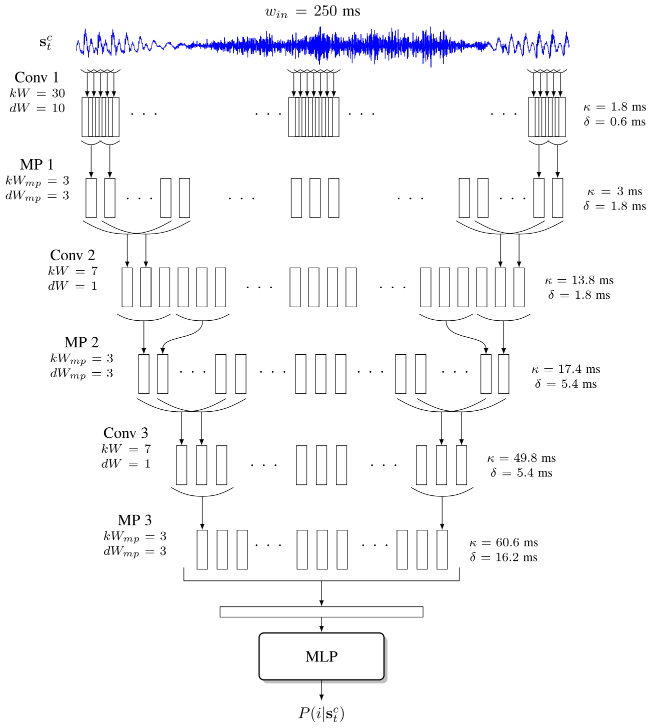


Fig. 7. Illustration of the feature stage of CNN trained on TIMIT to classify 183 phoneme classes. κ and δ indicates the temporal information modeled by the layer and the shift respectively. Non-linearity layers are applied after each max-pooling.

ANN-1H has 1000 nodes in the hidden layer and ANN-3H has 1000 nodes in each hidden layer.

- MediaParl: We use the setup of the best performing hybrid HMM/ANN system using a three hidden layers MLP, classifying 1101 and 1084 clustered context-dependent units for Swiss German and Swiss French respectively, reported in Razavi et al. (2014) and in Razavi and Magimai.-Doss (2014) as the baseline ANN-3H system. The ANN-1H has 1000 nodes in each hidden layer. The ANN-3H has 1800 nodes in the first hidden layer and 1500 nodes in the second and third hidden layer. The inputs to the ANNs were 39 PLP cepstral features ($c_0 - c_{12} + \Delta + \Delta\Delta$) with four frames preceding and four frames following context. The frame size and frame shift are 25 ms and 10ms, respectively.

4.2.2. Proposed CNN-based system

We train the proposed CNN-based $P(i|\mathbf{s}_t^c)$ estimator using raw speech signal. The inputs are simply composed of a window of the speech signal (hence $d_{in} = 1$, for the first convolutional layer). The utterances are normalized such that they have zero mean and unit variance, which is in line with the literature (Sheikhzadeh and Deng, 1994). No further pre-processing is performed. The hyper-parameters of the network are: the time span of the input signal (w_{in}), the kernel width kW and shift dW of the convolutions, the number of filters d_{out}, maxpooling kernel width kW_{mp}, maxpooling kernel shift dW_{mp} and the number of nodes in the hidden layer(s). Note that the input d_{in} for the first convolution layer is one (i.e. a sample of the speech signal). For the remaining layers, the d_{in} is the product of d_{out} of the previous layer and kW of that layer. These hyper parameters are determined by early stopping on the validation set, based on frame classification accuracy. The ranges which are considered for a coarse grid search are reported in Table 1. We use the TIMIT task to narrow down the hyper-parameters search space, as it provided fast turnaround experiments.

Table 1
Ranges of hyper parameters for the grid search.

Parameters	Units	Range
Input window size (w_{in})	ms	100–700
Kernel width of the first conv. (kW_1)	Samples	10–90
Kernel width of the n^{th} conv. (kW_n)	Frames	1–11
Number of filters per kernel (d_{out})	Filters	20–100
Max-pooling kernel width (kW_{mp})	Frames	2–6
Number of hidden units in the classifier	Units	200–1500

Table 2
Number of samples processed per second for the baselines and the proposed approach, during the training and evaluation phases. The measurements were done on a single CPU Intel i7 2600K 3.4 GHz.

System	Training [*sample/sec*]	Evaluation [*sample/sec*]
ANN-1H	1371	3330
ANN-3H	177	2199
CNN-1H	240	1164
CNN-3H	113	741

Table 3
Architecture of CNN-based system for different tasks. HL = 1 denotes CNN-1H and HL = 3 denotes CNN-3H. w_{in} is expressed in terms of milliseconds. The hyper-parameters kW, dW, d_{out} and kW_{mp} for each convolution layer is comma separated. HU denotes the number of hidden units. 2 × 1500 means 1500 hidden units per hidden layer.

	HL	w_{in}	kW	dW	d_{out}	kW_{mp}	HU
TIMIT	1	250	30,7,7	10,1,1	80,60,60	3,3,3	1000
	3	250	30,7,7	10,1,1	80,60,60	3,3,3	3x1000
WSJ	1	210	30,7,7	10,1,1	80,60,60	3,3,3	1000
	3	310	30,7,7	10,1,1	80,60,60	3,3,3	3x1000
MP-DE	1	210	30,7,7	10,1,1	80,60,60	3,3,3	1000
	3	310	30,7,7	10,1,1	80,60,60	3,3,3	1800,2x1500
MP-FR	1	190	30,7,7	10,1,1	80,60,60	3,3,3	1000
	3	310	30,7,7	10,1,1	80,60,60	3,3,3	1800,2x1500

For each of the tasks, we train CNNs with one hidden layer (denoted as CNN-1H) and three hidden layers (denoted as CNN-3H) similar to the different MLP architectures in the baseline systems. We found that three convolution layers consistently yield the best cross validation accuracy across all the tasks. The CNN architecture found for each of the task is presented in Table 3. The shift of max-pooling kernel $dW_{mp} = 3$ is found for all the layers on all the tasks. As we will observe later, the complexity of the CNN-based approach in terms of number of parameters lies at the classifier stage. So, for fair comparison with the baseline systems, we restricted the search for the number of hidden nodes in the hidden layer(s) such that the number of parameters is comparable to the respective baseline systems. The output classes are the same as the case of cepstral feature-based system, i.e. for the TIMIT task 183 phone classes, for the WSJ task 2776 cCD units, for the MP-DE task 1101 cCD units and for the MP-FR task 1084 cCD units.

The computation cost of the proposed architecture would be higher than the ANN baseline, as the raw speech signal has to be processed, whereas for the baseline systems the features are already computed. Table 2 presents the number of frames processed per second for the baseline, the CNN-1H and the CNN-3H systems during the training and evaluation phases. One can see that while training the baseline with one hidden layer (ANN-1H) is much faster than training the CNN-1H (5.7x speed factor), the gap reduces drastically when training the three layers systems (1.5x speed factor).

4.3. Results

In this section we present the results of the studies on different tasks. For the sake of completeness, for the speech recognition studies we also report performance on HMM/GMM system. For MP-

Table 4
Phoneme error rate of different systems on the core test set of the TIMIT corpus. The ANN-1H and ANN-3H performances are reported in Mohamed et al. (2012). #Conv. Params. denotes the number of parameters in the convolution layers, #Class. Params. denotes the number of parameters in the classifier stage. M stands for million.

Input	System	#Conv. params.	#Class. params.	PER (in %)
MFCC	ANN-1H	na	1.2M	24.5
MFCC	ANN-3H	na	2.6M	22.6
RAW	CNN-1H	63k	0.92M	22.8
RAW	CNN-3H	52k	2.9M	21.9

Table 5
Phoneme error rate of different systems reported in literature on the core test set of the TIMIT corpus.

Method (input)	PER (in %)
Augmented CRFs (MFCC) (Hifny and Renals, 2009)	26.6
HMM/DNNs 6 layers (MFCC) (Mohamed et al., 2012)	22.3
Deep segmental NN (MFCC) (Abdel-Hamid et al., 2013)	21.9
Proposed approach	**21.9**
HMM/DNNs 6 layers (MFCC+LDA+MLLT+fMLLR) (Lu et al., 2016)	18.5
CTC transducers (FBANKs) (Graves et al., 2013)	17.7
Attention-based RNN (FBANKs) (Chorowski et al., 2015)	17.6
Segmental RNN (MFCC+LDA+MLLT+fMLLR) (Lu et al., 2016)	17.3

DE and MP-FR, the best performing HMM/GMM systems reported in Razavi et al. (2014) and Razavi and Magimai.-Doss (2014), respectively are presented. These systems have a greater number of tied states than the hybrid HMM/ANN and the CNN-based system presented here.

4.3.1. TIMIT

Table 4 presents the results on TIMIT phone recognition task in terms of phoneme error rate (PER). It can be observed that the proposed CNN based approach outperforms the conventional cepstral feature-based system. In Mohamed et al. (2012, Fig. 6), ANNs with different hidden layers were investigated with cepstral feature as input. The best performance of 23.0% PER for the case of random initialization is achieved with 7 hidden layers, 3072 hidden nodes per layer and 17 frames temporal context (8 preceding and 8 following). With pre-training, the best performance of 22.3% PER is achieved with 6 hidden layers, 3072 hidden nodes per layer and 17 frames temporal context. The CNN-3H system performs better than those systems as well.

Table 5 contrasts our results with a few prominent results on TIMIT using ANNs. Inputs of these systems are either MFCCs (computed as presented in Section 4.2.1), Mel filterbanks energies (abbreviated as FBANKs) or “improved” MFCC features (denoted as MFCC+LDA+MLLT+fMLLR), which are obtained by applying decorrelation processes (linear discriminant analysis and maximum likelihood linear transform) and speaker normalization (feature-space maximum likelihood linear regression) (Rath et al., 2013) to the original MFCC coefficient. One can see that the proposed approach outperforms most of the systems using MFCCs features. Systems using improved MFCCs features yields better results than the proposed approach, mainly due to the speaker normalization technique, which could be developed for the proposed approach. For instance, speaker adaptation in our approach could be achieved in an unsupervised manner by using learning hidden unit contributions (LHUC) method at the classifier/MLP stage (Swietojanski et al., 2016). At the filter stage, one could possibly adopt an approach similar to the approach proposed in (Abdel-Hamid and Jiang, 2013). Finally, one can see that RNN-based systems (the three last entries of Table 5) clearly yield the best performance. It is worth noting that the proposed CNN-based approach could be used in a RNN-based architecture, where the MLP-based classifier stage is replaced by a RNN. This approach raises the issue of the high dimensionality of the filter stage output. It could be addressed by adding more convolution and max-pooling layers, which will effectively reduce the output dimensionality. Such an approach we have explored successfully in the context of extension of our approach where the MLP is replaced by single layer perceptron to reduce the overall complexity of the system in terms of parameters while retaining the performance (Palaz et al., 2014).

Table 6
Word Error Rate on the Nov’92 testset of the WSJ corpus. #Conv. Params. denotes the number of parameters in the convolution layers, #Class. Params. denotes the number of parameters in the classifier stage. M stands for million.

Input	System	#Conv. params.	#Class. params.	WER (in %)
MFCC	GMM	na	4M	5.1
MFCC	ANN-1H	na	3.1M	7.0
MFCC	ANN-3H	na	5.6M	6.4
RAW	CNN-1H	46k	3.1M	6.7
RAW	CNN-3H	61k	5.6M	5.6

Table 7
Word Error Rate on the testset of the MP-DE corpus. The GMM and ANN-3H baseline performances are reported in Razavi et al. (2014). #Conv. Params. denotes the number of parameters in the convolution layers, #Class. Params. denotes the number of parameters in the classifier stage. M stands for million.

Input	System	#Conv. params.	#Class. params.	WER (in %)
PLP	GMM	na	3.8M	26.6
PLP	ANN-1H	na	2.2M	26.7
PLP	ANN-3H	na	8.8M	25.5
RAW	CNN-1H	61k	1.6M	24.4
RAW	CNN-3H	92k	8.7M	23.5

Table 8
Word Error Rate on the testset of the MP-FR corpus. The GMM and ANN-3H performances are reported in Razavi and Magimai.-Doss (2014). #Conv. Params. denotes the number of parameters in the convolution layers, #Class. Params. denotes the number of parameters in the classifier stage. M stands for million.

Input	System	#Conv. params.	#Class. params.	WER (in %)
PLP	GMM	na	3.8M	26.8
PLP	ANN-1H	na	2.2M	27.0
PLP	ANN-3H	na	8.8M	25.5
RAW	CNN-1H	61k	1.5M	25.9
RAW	CNN-3H	92k	8.7M	23.9

4.3.2. WSJ

The results for the LVCSR study (Palaz et al., 2015b) on the WSJ corpus is presented in Table 6. For the baseline systems and the proposed system. As can be observed, the CNN-1H based system outperforms the ANN-1H based baseline system, and the CNN-3H based system also outperforms the ANN-3H based system with as many parameters.

4.3.3. MP-DE

The results on the Mediaparl German corpus are presented in Table 7. The CNN-1H based system outperforms the GMM-based system, the ANN-1H based system and the ANN-3H system with four times less parameters. The CNN-3H system also outperforms the baseline.

4.3.4. MP-FR

The results on the Mediaparl French corpus are presented in Table 8. Again, a similar trend can be observed, i.e. the CNN-1H based system outperforms the ANN-1H baseline and the CNN-3H outperforms the ANN-3H based system.

In summary, these studies show that with minimal assumptions the proposed approach is able to learn to process the speech signal to estimate phone class conditional probabilities $P(i|\mathbf{s}_t^c)$ and yield a system that outperforms conventional cepstral feature based system using DNNs. Furthermore, we consistently observe that the CNN-1H system yields performance comparable to ANN-3H system with considerably fewer parameters.

5. Analysis

The aim of this section is to gain insight into the proposed approach. Towards that this section focuses on analysis at two levels: (a) analysis of the first convolution layer (Section 5.1) which operates on the speech signal directly. Thus, can be related to and can be contrasted with traditional speech processing; and (b) analysis of the intermediate feature representations obtained at the output of the feature stage (Section 5.2).

5.1. First convolution layer

In this section, we present an analysis of the first convolution layer. We first provide an input level analysis, where the hyper-parameters of the layer (found experimentally) are compared against the conventional speech processing approach. We then show that the convolution layer can be interpreted as a bank of matching filters. Finally, we analyze how these filters respond to various inputs and present a method to understand the filtering process.

5.1.1. Input level analysis

To learn to process raw speech signal and estimate $P(i|\mathbf{s}_t^c)$ the proposed approach employs many hyper-parameters which are decided based on validation data. We can get insight into the approach by relating or contrasting a few of the hyper-parameters to the traditional speech processing. First among that is time span of the signal w_{in} used to estimate $P(i|\mathbf{s}_t^c)$. From Table 3, we can observe that w_{in} varies from 190 ms to 310 ms. This is consistent with the literature which supports the idea of processing syllable length speech signal (around 200 ms) for classification of phones (Hermansky, 1998). This aspect can be also observed in another way. Usually, in hybrid HMM/ANN system the input is the cepstral features (static + Δ + ΔΔ) at the current time frame and features of four preceding frames and four following frames. If the frame shift is 10 ms and the temporal derivatives are computed using two frames preceding and two frames following context then the 9 frame feature input models 170 ms of speech signal.

Next, we can understand how the speech signal of time span of 190–310 ms is processed at the input of the network through the kernel width (kW) and kernel shift (dW) of the first convolution stage. We can see from Table 3 that for all tasks kW is 30 speech samples and dW is 10 speech samples. Given that the sampling frequency is 16 kHz, this translates into a window of 1.8 ms and shift of about 0.6 ms. This is contrary to the conventional speech processing where typically the window size is about 25 ms, the shift is about 10 ms and the resulting features are concatenated at the classifier input. Note that in our case w_{in} is shifted by 10 ms, however within the window of 190–310 ms the speech is processed at the sub-segmental level at the first convolution layer and subsequently processed by later convolution layers to estimate $P(i|\mathbf{s}_t^c)$.

Such a sub-segmental processing at the first convolution layer could possibly be reasoned through signal stationarity assumptions. More precisely, the convolution filters at the first stage are learned by discriminating the phone classes at the output of the CNN. So, for the output of the convolution filter to be informative (for phone classification), the filter has to operate on stationary segments of the speech signal spanned by w_{in}. It can be argued that such a stationary assumption would clearly hold for one glottal cycle or pitch period of the speech signal. In such a case suppose if the limit of the observed pitch frequency is assumed to be 500 Hz, i.e. beyond adult speakers' pitch frequency range, then a window size of 2 ms or less would ensure that the filters operate on stationary segments, i.e. within a glottal cycle, which mainly contains vocal tract response related information. This is consistent with traditional feature extraction methods (see Rabiner and Juang (1993), Davis and Mermelstein (1980) and Hermansky (1990) for example), where the main emphasis is towards modeling vocal tract response information.

5.1.2. Learned filters

The first convolution layer learns a set of filters that operates on the speech signal in a similar way to filter bank analysis during MFCC or PLP cepstral feature extraction. In the case of MFCC or PLP cepstral feature extraction the number of filter banks and their characteristics are determined a priori using speech perception knowledge. For instance, the filters are placed either on Mel scale or on Bark scale. Further, each of the filters covers only a part of the bandwidth, out of which the response is strictly zero. The number of filters is chosen based on bandwidth information. For instance, in the case of Mel scale around 24 filters for 4 kHz bandwidth (narrow band speech) and 40 filters for 8 kHz bandwidth (wide band speech) are typically used. While in the case of Bark scale, there are 15 filters for 4 kHz bandwidth and 19 filters for 8 kHz bandwidth (Hönig et al., 2005).

In contrast, in the proposed approach the number filters and their responses are learned in data-driven manner, i.e. while learning to estimate $P(i|\mathbf{s}_t^c)$. It can be observed from Table 3 that the number of filters for all the tasks is 80. This is well above the range typically used in speech processing. In order to understand the learned filter characteristics, we analyzed the filters learned on WSJ, MP-DE and MP-FR task in the following manner:

(i) The complex Fourier transform $\mathcal{F}$ of the filters learned on the WSJ, MP-DE and MP-FR tasks for CNN-1H case are computed using 1024 point FFT. The 512 point magnitude spectrum $|\mathcal{F}_m|$ of each filter m is then normalized, i.e. converted into a probability mass function. F_m denotes the normalized magnitude spectrum of filter m.

(ii) For each filter $m = 1, \dots, 80$ learned on WSJ, we find the closest filter $n = 1, \dots, 80$ learned on MP-DE and MP-FR using symmetric Kullback-Leibler divergence,

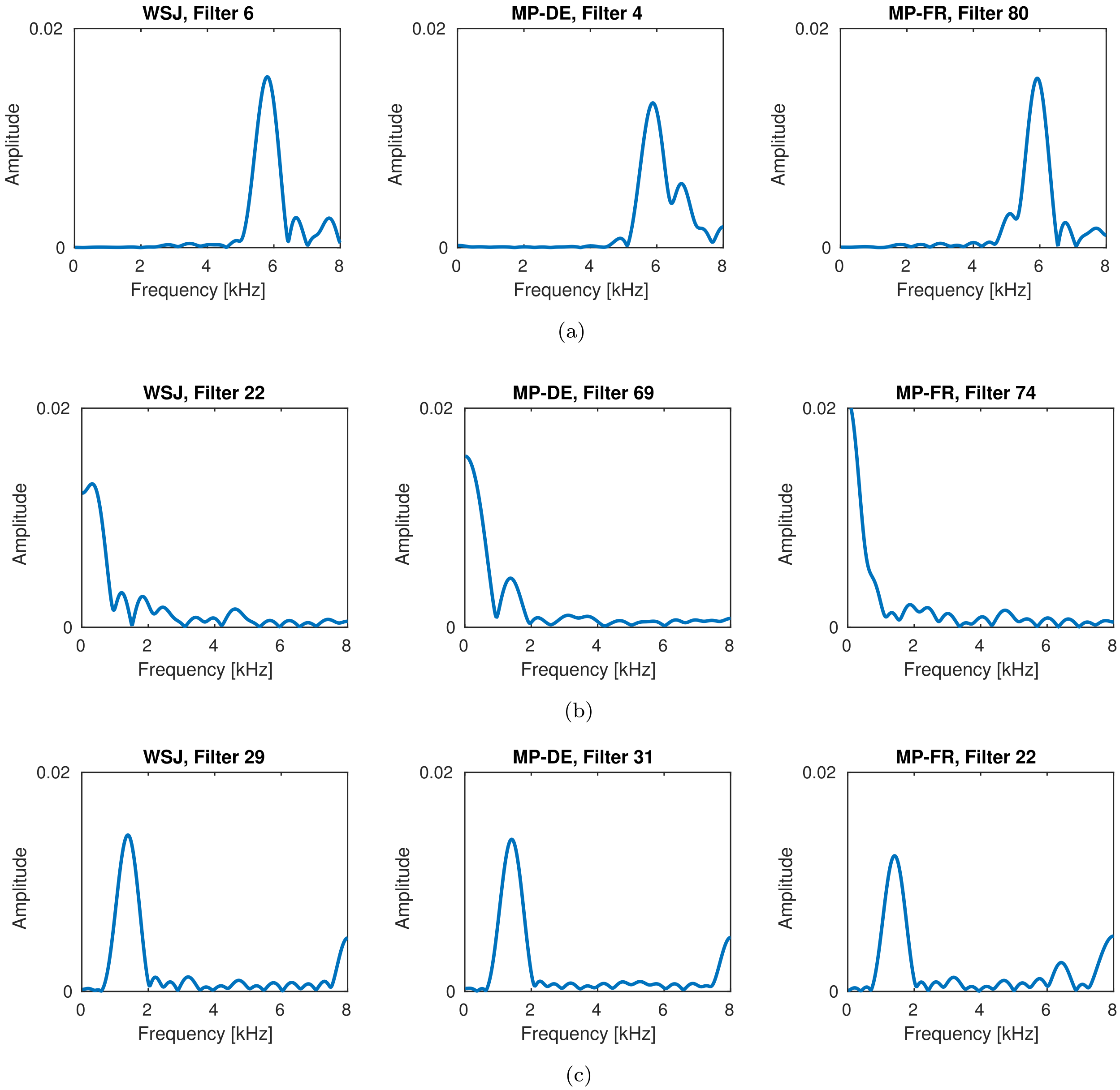
$$d(F_m, F_n) = \frac{1}{2} \cdot [D_{KL}(F_m||F_n) + D_{KL}(F_n||F_m)], \tag{9}$$

$$D_{KL}(F_m||F_n) = \sum_{u=1}^{512} F_m^u \ln \frac{F_m^u}{F_n^u}, \tag{10}$$

where F_m^u is the normalized magnitude at uth point of FFT of filter m of WSJ CNN-1H and F_n^u is the normalized magnitude at uth point of FFT of filter n of MP-DE CNN-1H or MP-FR CNN-1H.

Fig. 8 presents the magnitude of the Fourier transform of a few filters learned on WSJ (on the left column) and the closest filters learned on the MP-DE task (on the middle column) and on the MP-FR task (on the right column). We can make two observations. First, the filters are focusing on different parts of the spectrum. However, unlike the filter banks in the MFCC or PLP cepstral feature extraction, the frequency response of the filters covers the whole bandwidth. Second, it can be observed that similar filters can be found across domains and languages, although there is a difference in the spectral balance, especially as observed in the case of Fig. 8(b).

To further visualize the learned filters, we ordered the filters according to the frequency at which the response is maximum. We treat these frequencies as the center frequencies of the learned filters. Fig. 9 plots the center frequencies of the learned filters along with the center frequencies of 80 critical bands mel-scale filter bank and Gammatone filter bank. It can be observed that the learned filter placements to a certain extent tend to match the auditory motivated filter banks, in particular



(a)

(b)

(c)

Fig. 8. Examples of three close pairs of filters learned. The left column is from CNN-1H WSJ, the center one is from CNN-1H MP-DE, the right one is from CNN-1H MP-FR.

mel scale filter bank, in the lower half of the bandwidth, i.e. 0 Hz and 4 kHz but differ considerably in the upper half of the bandwidth. In contrast, in the works of Sainath et al. (2015) and Tüske et al. (2014) the filter placements were found to be close to Gammatone filter bank. A potential reason for this difference could be that in these works the filter lengths and the number of filters were set based on prior knowledge. When comparing across WSJ, MP-DE and MP-FR, the learned filter placements for MP-FR and MP-DE are similar to each other but differ from that of WSJ. Having said that it is worth pointing out that the learned filters can have more than one pass band, as can be seen in Fig. 8. So generalizing these observations in comparison to auditory motivated filter banks is not trivial.

To further understand the characteristics of the learned filters, we estimated the cumulative frequency response of all the learned filters:

$$F_{cum} = \sum_{n=1}^{80} F_n \tag{11}$$

Fig. 10 presents the gain normalized cumulative frequency responses for CNN-1H WSJ, CNN-1H MP-DE and CNN-1H MP-FR. We can make three key observations,

(i) Emphasis is given to frequency regions below 3500 Hz (telephone bandwidth) and high frequency region in the range of 6000–8000 Hz.
(ii) Though the filters are learned on different languages and corpora, we can see that below 4000 Hz and above 6500 Hz the frequency response for WSJ, MP-DE and MP-FR are similar. As the filters are operating on sub-segmental speech, we speculate that the peaks (high energy regions) are more related to the resonances in the vocal tract or phoneme discriminative invariant information. Between 4000 Hz and 6500 Hz, we can see that MP-DE and MP-FR have responses that closely match but are different than WSJ. Overall, we observe that the spectral balance for WSJ is different than for MP-DE and MP-FR. We attribute this balance mismatch mainly to the fact that the WSJ and the Mediaparl corpora are

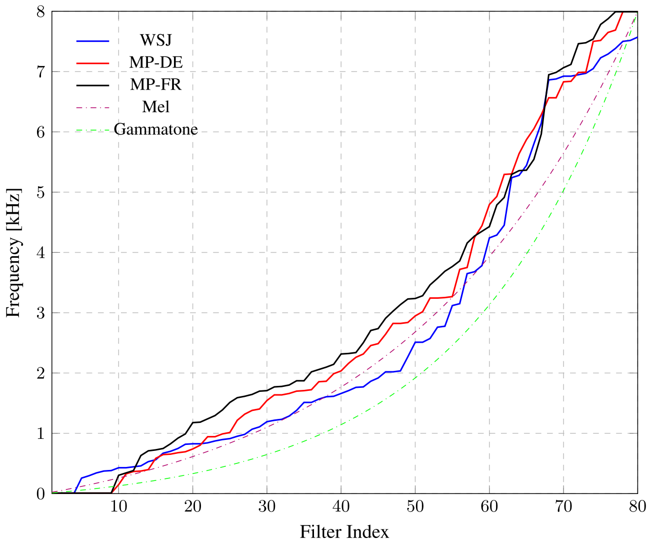


Fig. 9. Plot of learned filters for WSJ, MP-DE and MP-FR ordered according to the frequency of maximum response along with the center frequencies of 80 critical band Mel-scale and Gammatone filter banks.

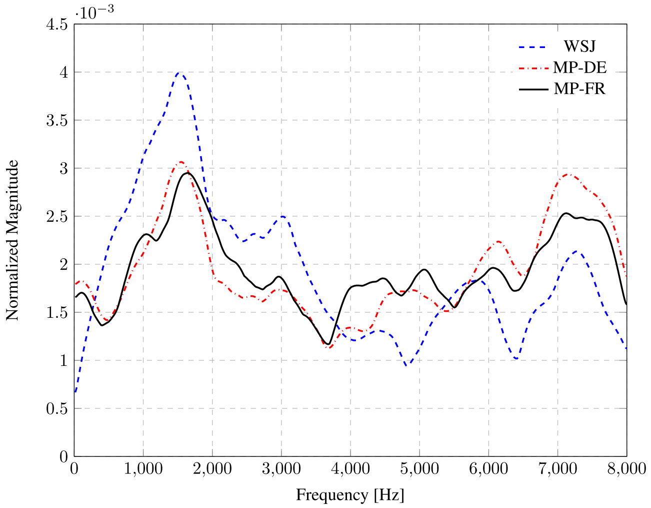


Fig. 10. Cumulative frequency responses of the learned filters on WSJ, MP-DE and MP-FR.

different domains in terms of type of speech (read vs. spontaneous) and recording environment (controlled vs real world). In the following sub-section and Section 5.2.2 we touch upon this aspect again.

(iii) Auditory filter banks such as Mel scale filter banks or Bark scale filter banks are usually designed to have a cumulative frequency response that is flat. In other words, constant Q bandpass filter bank. In contrast to that, it can be seen that the cumulative frequency response of the learned filters is not constant Q bandpass. The main reason for that is standard filter banks emerged from human sound perception studies considering the complete auditory frequency range or the bandwidth, so as to aid analysis and synthesis (reconstruction) of the audio signal. However, in our case these filters are learned for the purpose of discriminating phones, and the speech signal contains information other than just phones. The figure suggests that, for discriminating only phones, constant Q bandpass filter bank is not a necessary condition.

5.1.3. *Response of filters to input speech signal*

In Section 5.1.1, we observed that the speech signal of time span 190–310 ms is processed in sub-segmental manner. In the previous section, we observed that the filters that operate on sub-segment of speech signal are tuned to different parts of the spectrum during training. In other words, matched to different parts of the spectrum relevant for phone discrimination. In this section, we ascertain that by analyzing the response of the filters to the input speech signal in relationship with phones.

The CNNs in the WSJ, MP-DE and MP-FR studies are trained to classify cCD units, which can be quite distinctive across languages. So, in order to facilitate the analysis across languages, we train CNNs with single hidden layer on WSJ, MP-DE and MP-FR data to classify context-independent phones with the same hyper parameters. We denote these CNNs as CNN-1H-mono WSJ, CNN-1H-mono MP-DE and CNN-mono MP-FR, respectively.

As a first step, we analyze the energy output of the filters to the input speech signal. Formally, for a given input $\mathbf{s}_t = \{s_{t-(kW-1)/2} \dots s_{t+(kW-1)/2}\}$, the output $\mathbf{y}_t$ of the first convolution layer is given by:

$$\mathbf{y}_t[m] = \sum_{l=-(kW-1)/2}^{l=+(kW-1)/2} f_m[l] \cdot s_{t+l} \quad \forall m = 1,.,d_{out} \tag{12}$$

where f_m denotes the mth filter in first convolution layer and $\mathbf{y}_t[m]$ denotes the output of the filter at time frame t. Fig. 11 presents the output of the filters of CNN-1H-mono WSJ given a segment of speech signal corresponding to phoneme /I/ as input.

It can be seen that at each time frame only a few filters out of the 80 filters have high energy output. An informal analysis across different phones showed similar trends, except that the filters with high energy output were different for different phones. Together with the findings of the previous section, this suggests that the learned filters could be a *dictionary* that models the information in the frequency domain *in-parts* for each phone. With that assumption, we extended the analysis where,

1. the magnitude spectrum $\mathcal{S}_t$ of the input signal $\mathbf{s}_t$ based on the dictionary is estimated as:

$$\mathcal{S}_t = |\sum_{m=1}^{M} \mathbf{y}_t[m] \cdot \mathcal{F}_m|, \tag{13}$$

where $\mathbf{y}_t[m]$ is the output of filter m as in Eq. (12) and $\mathcal{F}_m$ is the complex Fourier transform of filter f_m.
It is worth noting that if the dictionary was to correspond to a bank of kW Fourier sine and cosine bases then $\mathcal{S}_t$ is nothing but the Fourier magnitude spectrum of the input signal $\mathbf{s}_t$. As $\mathbf{y}_t[m]$ would be a projection on to the Fourier basis corresponding to discrete frequency m, and $\mathcal{F}_m$ would *ideally* be a Dirac delta distribution centred at the discrete frequency m.
2. A frame level magnitude spectrum $\mathcal{S}_i$ for phone i is estimated by averaging the magnitude spectrum $\mathcal{S}_t$ obtained over speech signal of length equal to frameshift, which in our case is 10 ms. More precisely, with in 10ms speech, $\mathcal{S}_t$ is estimated every 10 samples as per Eq. (13) and averaged by the number of sub-segmental frames in 10 ms or 160 samples speech, i.e. 16. $\mathcal{S}_i$ can be seen as the average spectral information that is modeled every 10 ms.

We performed a qualitative analysis on American English vowels dataset, which contains 12 vowels produced by 45 men, 48 women, and 46 in h-V-d syllables (e.g., had, hid, hood) (Hillenbrand et al., 1995). The analysis was carried out using the filters in the first convolution layer of WSJ CNN-1H-mono. We used 256 points for DFT. Fig. 12 presents the $\mathcal{S}_i$ estimated for a frame of /ah/, /eh/, /er/, /oa/, /uw/ and /iy/ produced by male speaker m01, female speaker w01, boy speaker b01 and girl speaker g01. In the plots, the observed first and second spectral peaks have been marked and contrasted with the F1-F2 (first format-second formant) range obtained in the coarse sampling part of the orig-

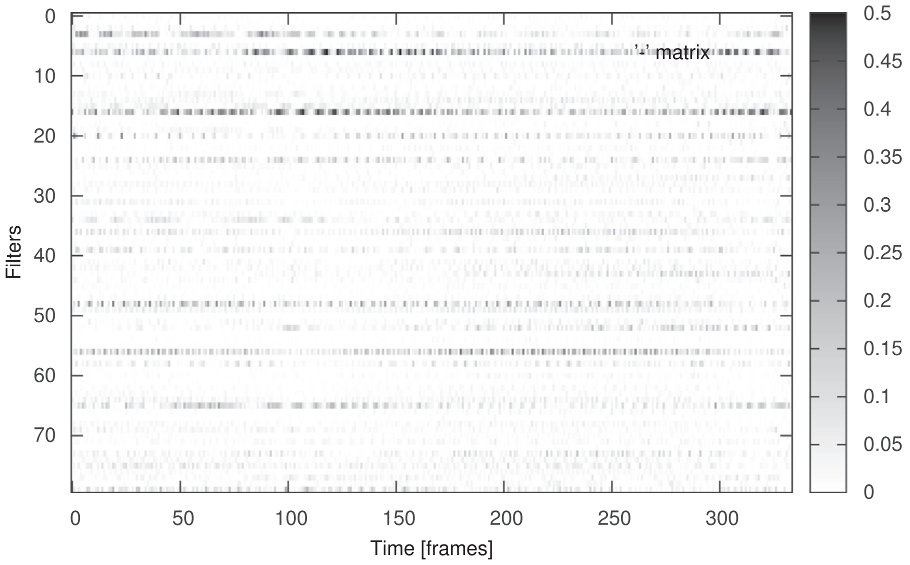


Fig. 11. Normalized energy output of each filter in the first convolution layer of CNN-1H-mono WSJ for an input speech segment corresponding to phoneme /I/.

inal study by Hillenbrandt et al.[2] It can be observed that the spectrum estimates are different for different vowels. Furthermore, except for few cases, the marked spectral peaks correspond to F1-F2 range. In the case of /ah/, only one peak is discernible due to merger of the first two formants. Similarly, in the case of /oa/ for speaker b01, only one peak is discernible due to merger of the first two formants. Merging of formants appears to happen in the case for the second spectral peak of /iy/ of speaker g01, due to merger of F2 and F3. The magnitude spectrum also has ripples. The ripples and the merger of close by formants could potentially be a consequence of the short kernel width i.e. 30 samples, i.e. sub-segmental speech processing. We performed similar analysis on a few other speakers in the American vowel dataset and found that the detected peaks tend to correspond to F1-F2 formant ranges obtained in the original acoustic analysis study. It is interesting to note that the analysis holds well for children speech, despite the net being trained on adult speech.

American English vowel dataset is a controlled dataset, where the phonetic context is restricted. In order to ascertain that the observations made above holds true irrespective of the phonetic context or speakers, we performed an analysis on the validation data of WSJ, MP-DE and MP-FR using the filters in the first convolution layer of respective CNN-1H-mono, where given the segmentation the frame level spectrum estimates S_i are averaged across all the speakers for each phone i. We denote the speaker averaged spectrum as $\bar{S}_i$. Fig. 13 displays $\bar{S}_i$ of a few prominent vowels (notated in the SAMPA format) for WSJ, MP-DE and MP-FR. It can be observed that the frame level magnitude spectrum averaged across speakers is different for each vowel. This difference is particularly observable in the frequency regions below 4000 Hz and in the frequency regions between 6000 Hz and 8000 Hz. We had earlier observed in Section 5.1.2 that these are frequency regions that the learned filters give emphasis to. The prominent spectral peaks could be related to the formants. However, a detailed formant analysis similar to the frame level analysis on American English vowels dataset is practically infeasible for two main reasons:

(a) First, the formant frequencies and their bandwidths for males and females are different. The frequency responses here are result of averaging over several male and female speakers in the respective validation data set; and

(b) Second, the analysis here has been carried on validation data, not on actual training data. So there can be spurious information present due to unseen condition or variation.

For instance, in the case of /A/, see Fig. 13(e), we observe a prominent peak at around 1000 Hz, which could be seen as merger of first formant and second formant as a consequence of window effect and averaging over male and female speakers. Taking these aspects into account, we examined the frequency responses in the case of WSJ (Fig. 13(a)). We found that the prominent spectral peak locations tend to relate well to the first formant, second formant and third formant information provided for English vowels in Deng and O'Shaughnessy (2003, p. 233). It is worth mentioning that a similar observation that filters capture formant information has been made when learning jointly feature and classifier from short-term magnitude spectrum (Biem et al., 2001). When comparing across the languages (Fig. 13(d) and Fig. 13(e)) we observe a trend similar to the cumulative response of the filters (Fig. 10). Specifically, the spectral peak locations and spectral balance match well for MP-DE and MP-FR. However, in the case of WSJ the spectral peak locations tend to match but the spectral balance is different than MP-DE and MP-FR.

The analysis on American English vowels dataset, WSJ, MP-DE and MP-FR together indicates that the first convolution layer is learning formants related information.

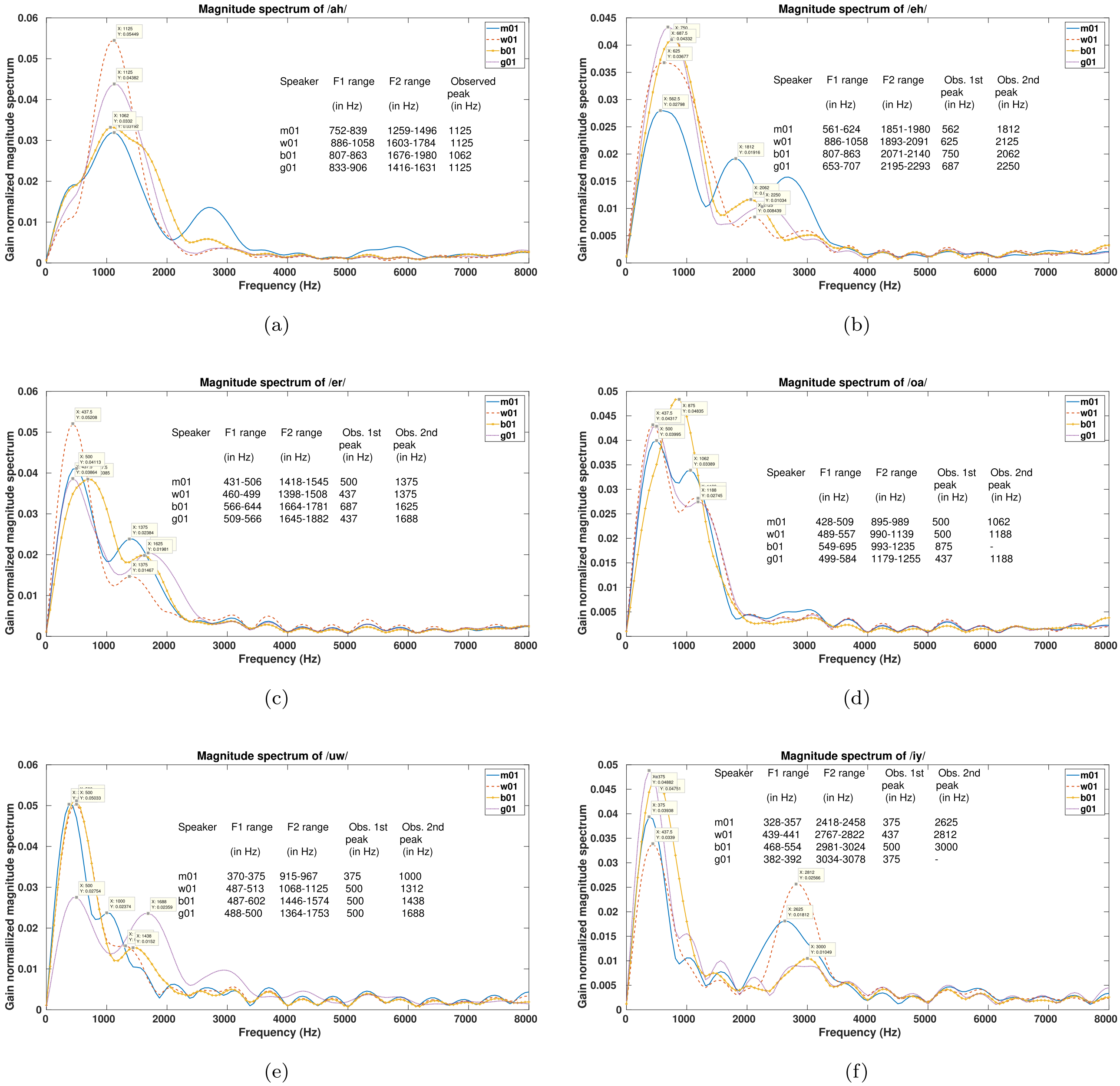
5.2. Intermediate feature level analysis

In this section, we focus on the analysis of intermediate feature representations that are being learned at the output of the feature learning stage. In that regard, Section 5.2.1 focuses on the discriminative aspects of the learned feature representations. Section 5.2.2 then focuses on the cross-domain and cross-lingual aspects.

5.2.1. Discriminative features

In the recognition studies presented earlier in Section 4, it was observed that CNN-1H system with much fewer parameters outperforms ANN-3H system on all the tasks. Furthermore, we also observed that the complexity of the proposed CNN-based system lies more at the classifier stage. Given that the intermediate feature representations are learned in the process of training $P(i|\mathbf{s}_t^c)$ estimator, it can be presumed that these features are more discriminative compared to cepstral-based feature representations, and thus needs less parameters at the classifier stage. To fully ascertain that aspect, we conduct an experiment to compare the cepstral features and the intermediate feature representations learned by the CNN. Specifically, we train and test three single layer perceptron

[2] https://homepages.wmich.edu/~hillenbr/voweldata/vowdata.dat.



Speaker	F1 range (in Hz)	F2 range (in Hz)	Observed peak (in Hz)
m01	752-839	1259-1496	1125
w01	886-1058	1603-1784	1125
b01	807-863	1676-1980	1062
g01	833-906	1416-1631	1125

(a)

Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	561-624	1851-1980	562	1812
w01	886-1058	1893-2091	625	2125
b01	807-863	2071-2140	750	2062
g01	653-707	2195-2293	687	2250

(b)

Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	431-506	1418-1545	500	1375
w01	460-499	1398-1508	437	1375
b01	566-644	1664-1781	687	1625
g01	509-566	1645-1882	437	1688

(c)

Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	428-509	895-989	500	1062
w01	489-557	990-1139	500	1188
b01	549-695	993-1235	875	-
g01	499-584	1179-1255	437	1188

(d)

Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	370-375	915-967	375	1000
w01	487-513	1068-1125	500	1312
b01	487-602	1446-1574	500	1438
g01	488-500	1364-1753	500	1688

(e)

Speaker	F1 range (in Hz)	F2 range (in Hz)	Obs. 1st peak (in Hz)	Obs. 2nd peak (in Hz)
m01	328-357	2418-2458	375	2625
w01	439-441	2767-2822	437	2812
b01	468-554	2981-3024	500	3000
g01	382-392	3034-3078	375	-

(f)

Fig. 12. Magnitude spectrum S_i for a 10 ms frame of American English vowels (a) /ah/, (b) /eh/, (c) /er/, (d) /oa/, (e) /uw/ and (f) /iy/ of speakers m01, w01, b01 and g01. As mentioned earlier, the F1-F2 ranges were obtained from the coarse sampling part of the original study.

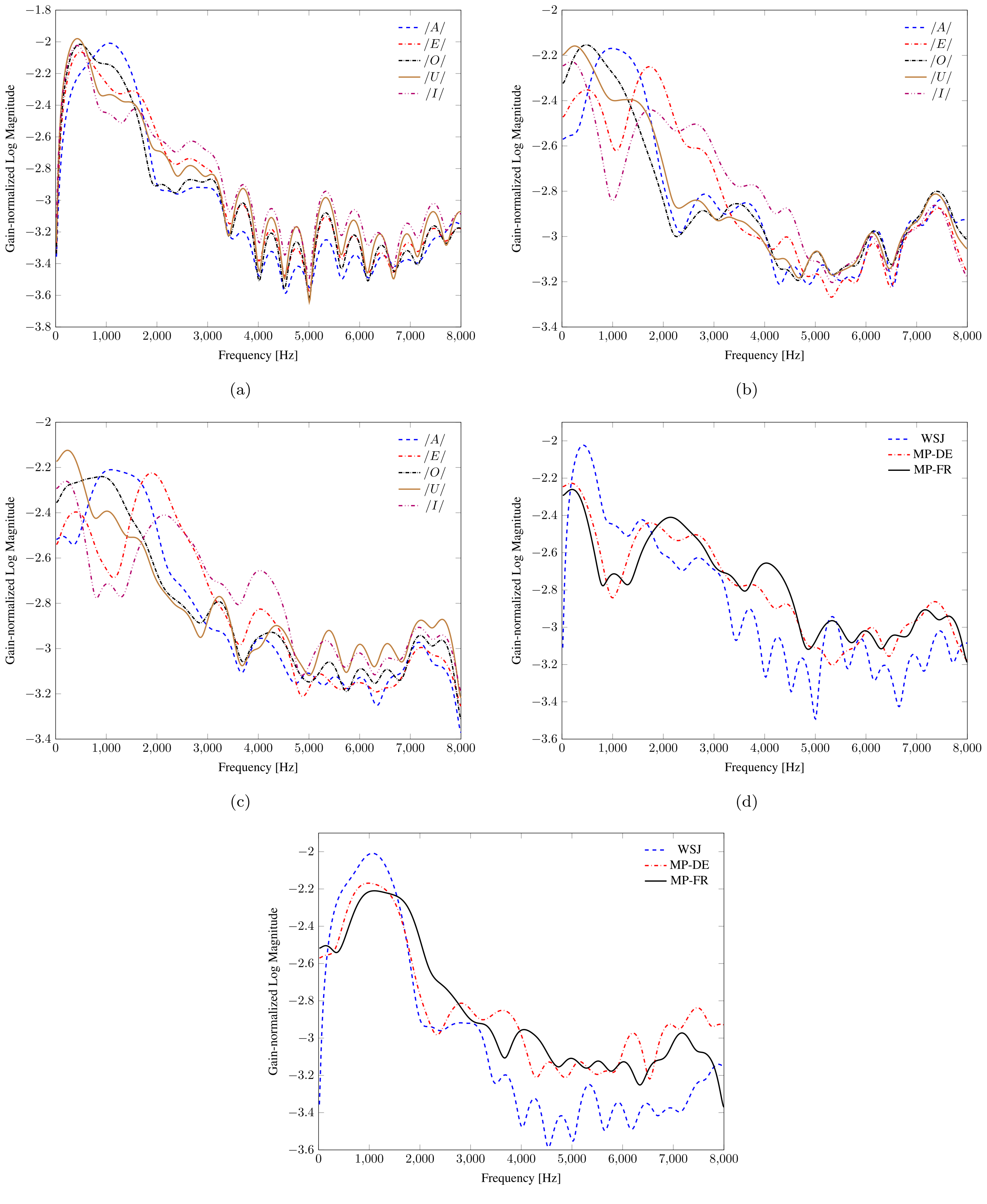
Table 9
Single layer perceptron-based system results on the Nov'92 test set of the WSJ task.

Features	Dimension	WER (in %)
MFCC	351	10.6
CNN-1H	540	7.9
CNN-3H	540	7.9

(SLP) based systems on WSJ task. One with the MFCCs with temporal context (39 × 9) as input and the others with intermediate features learned by CNN-1H and CNN-3H. In the case of CNN-3H, w_{in} is kept same as CNN-1H i.e. 210 ms. Table 9 presents the performances of the three systems. We can observe that the learned features lead to a better system than the cepstral features. Thus, indicating that the learned features are indeed more discriminative than the cepstral feature representation. Furthermore, it is interesting to note that the features learned by CNN-1H and CNN-3H yield similar systems. It suggests that the gain in ASR performance for the WSJ task using CNN-3H over CNN-1H is largely due to more hidden layers

5.2.2. *Cross-domain and cross-lingual studies*

Conventional cepstral-based features, like MFCC, are known to be independent of the language or the domain, which is one of the main reasons they become "standard" features. In the proposed system, the features are learned in a data-driven manner, thus they may have some level of dependencies on the data. In order to ascertain, to what extent the learned features are domain or language independent, we conducted cross-domain and cross-lingual experiments. More precisely, as illustrated in Fig. 14, in these experiments the filter stage was first trained on one domain or language. It was then used as feature extractor to train the classifier stage of another domain or language.



(a)

(b)

(c)

(d)

(e)

Fig. 13. Magnitude spectrum averaged across speakers $\bar{S}_i$ (a) for phonemes */E/*, */A/*, */O/*, */I/* and */U/* estimated by CNN-1H-mono WSJ; (b) for phonemes */E/*, */A/*, */O/*, */I/* and */U/* estimated by CNN-1H-mono MP-DE; (c) for phonemes */E/*, */A/*, */O/*, */I/* and */U/* estimated by CNN-1H-mono MP-FR; (d) for phoneme */I/* in WSJ, MP-DE and MP-FR; and (e) for phoneme */A/* in WSJ, MP-DE and MP-FR. The phonemes are notated in the SAMPA format.

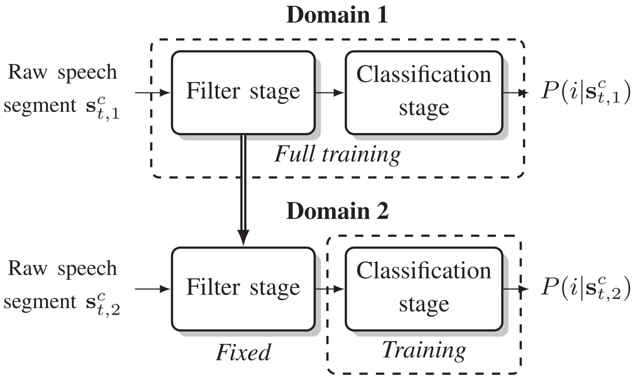

Domain 1
Raw speech segment $\mathbf{s}^c_{t,1}$
Filter stage
Classification stage
$P(i|\mathbf{s}^c_{t,1})$
Full training
Domain 2
Raw speech segment $\mathbf{s}^c_{t,2}$
Filter stage
Classification stage
$P(i|\mathbf{s}^c_{t,2})$
Fixed
Training

Fig. 14. Illustration of the cross-domain experiment. The filter stage is trained on domain 1, then used as feature extractor on domain 2.

Table 10
Cross-domain results on English. The TIMIT results are in terms of PER. The WSJ task results are in terms of WER.

Classifier stage (Domain 2)	Feature stage (Domain 1)	Error rate (in %)
TIMIT	Learned on TIMIT	22.8
	Learned on WSJ	23.3
WSJ	Learned on WSJ	6.7
	Learned on TIMIT	7.8

We use the TIMIT task and WSJ task for cross-domain experiments. We investigate

1. the use of feature stage of CNN-1H of WSJ task as feature extractor for the TIMIT task. The classifier stage with single hidden layer is trained on TIMIT to classify 183 phone classes.
2. the use of feature stage of CNN-1H of TIMIT task as feature extractor for the WSJ task. The classifier stage with single hidden layer is trained to classify 2776 clustered context-dependent units.

In both of the studies, we set the number of hidden nodes to 1000, similar to the systems reported in Section 4. The results of the two studies are presented in Table 10. In the case of TIMIT task the results are presented in terms of PER, and in the case of WSJ task in terms of WER. In the TIMIT task, we can observe that, despite the feature stage being trained to classify clustered context dependent units on a much larger corpus, the PER is inferior to the case where the feature stage is learned on TIMIT. In the case of WSJ task, we observe that with feature stage trained on TIMIT the WER is slightly worse (6.7% vs 7.8%).

In addition to the fact that TIMIT and WSJ are two different corpora, there are two other differences which could have had influence. First, WSJ is a much larger corpus than TIMIT in terms of data. Second, in TIMIT CNN-1H the feature stage is learned by classifying context-independent phones, while in WSJ CNN-1H the feature stage is learned by classifying clustered context-dependent units. So, we conducted a study on WSJ task to understand the influence of the type of units at the output of the CNN on the feature stage learning, while negating the data effect. More precisely, we use the feature stage of WSJ CNN-1H-mono (presented earlier in Section 5.1.3) as feature extractor and train the classifier stage to classify 2776 clustered context-dependent units. This system leads to a performance of 7.3% WER, which is inferior to 6.7% WER. This shows that indeed the type of units in the output of CNN has an influence on the feature learning stage. When compared to the case where the feature stage is learned on TIMIT, this result indicates that the performance gap is combined effect of the difference between the WSJ and TIMIT data sets and the units used at the output of the CNN learn the features. Finally, it is worth observing that TIMIT is a very small corpus compared to WSJ (3 h vs 88 h). However, the performance difference is not drastic, which suggests that the relevant features can be learned on relatively small amount of data.

Table 11
Cross-lingual studies result on English, German and French. The feature stage is learned on Domain 1 and the classifier stage is learned on Domain 2.

Classifier stage (Domain 2)	Feature stage (Domain 1)	WER (in %)
WSJ	Learned on WSJ	6.7
	Learned on MP-DE	12.1
	Learned on MP-FR	12.8
MP-DE	Learned on MP-DE	24.4
	Learned on MP-FR	26.1
	Learned on WSJ	30.9
MP-FR	Learned on MP-FR	25.9
	Learned on MP-DE	26.8
	Learned on WSJ	31.7

We investigate the cross-lingual aspects on WSJ, MP-DE and MP-FR tasks. We conduct studies where the feature stage is learned on one language and the classifier stage is learned on the other language. For these studies, we use the feature stages of WSJ CNN-1H, MP-DE CNN-1H and MP-FR CNN-1H systems presented in Section 4. The classifier stage in all the studies consisted of a single hidden layer with 1000 nodes. The classes at the output of classifier stage remained same as before, i.e. 2776 cCD units for the WSJ task, 1101 cCD units for the MP-DE task and 1084 cCD units for the MP-FR task. Table 11 presents the results of the study.

Before we analyze the results in detail, we can consider broader aspects. Specifically, in terms of family of languages, English and German belong to Germanic language family while French belongs to Romance language family. Given that, it can be expected that the feature stage learned on MP-DE to suit well for the WSJ task when compared to feature stage learned on MP-FR and vice versa. In the case of WSJ task this trend is observed (12.1% vs. 12.8%). However, it is not observed in the case of MP-DE task (30.9% vs. 26.1%). In general, we observe that feature stage learned on another language leads to inferior system. The performance gap is drastic when the feature stage is learned on WSJ and the classifier stage is learned on Medialparl (MP-DE or MP-FR) and vice versa. In addition to language differences, this can be attributed to the other differences in WSJ corpus and Medialparl corpus. More precisely, WSJ corpus contains read speech collected in controlled environment while Mediaparl contains spontaneous speech collected in real world conditions. This is also supported by the findings of the analysis presented in Section 5.1.2. Since MP-DE and MP-FR are similar kind of data except for the language, the drop in the performance is small (24.4–26.1% in the case of MP-DE task and 25.9–26.8% in the case of MP-FR task). Languages typically have different phone sets and this difference gets further enhanced when modeling context-dependent phones. As we saw earlier in the cross-domain studies the choice of output units influences the feature stage. So, the small drop in performance in this case can be more attributed to the phonetic level differences between German language and French language.

6. Discussion and conclusions

Motivated from recent advances in deep learning, the present paper investigated a novel CNN-based acoustic modeling approach that in a data- and task-driven manner determines the appropriate short-term processing, which consists of determining the window size and the number of filters for spectral processing, and learns the relevant representations from the speech signal to estimate phone class conditional probabilities for ASR. In this approach, the acoustic model consists of a feature stage and a classifier stage which are jointly learned during training. Specifically, the input to the acoustic model is raw speech signal, which is processed by several convolution layers (feature stage) and classified by an MLP (classifier stage) to estimate phone class conditional probabilities. We evaluated the approach against the conventional acoustic modeling approach, which consists of independent steps: short-term spectral based feature extraction and classifier training. Phone recognition studies on English and ASR studies on multiple languages (English, French, German) showed that the proposed acoustic modeling approach can yield better recognition systems.

To gain further insight, we performed analysis that largely focused on the filter stage of the approach. The key findings of the analysis are the following:

1. Both the conventional acoustic modeling approach and the proposed approach tend to model spectral information present in time span of about 200 ms for phone classification. However, they differ in the manner analysis is performed over that time span and feature representations are obtained. Indeed, in the proposed approach, contrary to the conventional wisdom of short-term processing, the signal is processed at the sub-segmental level (speech signal of about 2 ms) by the first convolution layer. The subsequent convolution layers temporally filter and integrate the output of first convolution layer to yield an intermediate representation. In other words, as illustrated in Fig. 7, the intermediate representation is obtained by processing the information at multiple temporal resolutions.
2. The filters in the first convolution layer learn from the sub-segmental speech a spectral dictionary that discriminate phones. Specifically, this dictionary was found to model formant related information. These findings are particularly interesting for different reasons. First, it validates the notion of formants and phone discrimination in a data-driven manner, i.e. without making an explicit assumption about speech production model. Secondly, sub-segmental spectral processing means high time resolution and low frequency resolution. Conventional method of short-term processing (i.e. determination of the window size) has been developed considering the trade-off between time resolution and frequency resolution and keeping analysis-synthesis in mind. Our investigations show that loss of frequency resolution due to sub-segmental speech processing is not affecting the ASR performance.
   Having said that, in Yegnanarayana and Veldhuis (1998), it has been shown that formant information can be effectively extracted through sub-segmental speech analysis. The method proposed in the above cited article considers details like closed and open glottal phases, positioning of the analysis window, choice of window size based on the gender information, choice of appropriate all-pole or pole-zero model to extract the formant information. The proposed approach does not make any such prior considerations while processing sub-segmental speech but still is found to model formant-like information. This could be indeed possible in our case without any such explicit considerations because, as pointed out in Section 5.1.1, the sub-segmental speech processed in the proposed approach is well below one pitch cycle of an adult male or female speaker (under normal speech conditions) and max pooling can provide shift invariance.
3. The intermediate feature representations learned at the output of the convolution stage are more discriminative than standard cepstral-based features. This reaffirms the point that learning the features and the classifiers jointly leads to more optimal systems when compared to conventional ”divide and conquer” approach.
4. The intermediate feature representations learned have some level of invariance across domains and languages. More specifically, in our analysis we observed that the variation of the learned features seems to come more from the domain characteristics as opposed to the set of subword units from the languages. This suggests that learning features in data-driven manner, as done using the proposed approach, could lead to language-independent features. This needs to be further investigated.

The proposed approach paves path for further research and development. We enumerate and discuss them briefly below.

1. noise robustness: as relevant features and classifier are automatically learned, a question that arises is: whether such an approach is robust in noisy conditions? In the analysis part, we have seen that the first convolution layer models envelop of sub-segmental speech signal spectrum. In particular formant-like information, which can be considered as high signal-to-noise ratio regions in the spectrum. Furthermore, subsequent processing through max pooling could be seen as filtering of spurious temporal information present in each filter output, while the second convolution layer filters could be interpreted along the lines of modeling envelop modulations in piecewise manner and combining them. Thus, the proposed approach could be expected to be robust. A preliminary investigation reported in Palaz et al. (2015a) and the investigations on Aurora2 and Aurora4 tasks reported in Palaz (2016, Chapter 5) indeed indicates that.
2. rapid adaptation of acoustic model: we have observed that the feature stage has considerably fewer parameters than the classifier stage. This provides new means to adapt the acoustic model. Specifically, one of the main challenges often faced in adapting the acoustic model to new domains is the amount of adaptation data available. The data may not be sufficient to effectively adapt all the parameters in the acoustic model. In the proposed approach, this challenge could be addressed by only adapting the feature stage. Such an approach would be analogous to maximum likelihood linear regression (MLLR) (Gales and Woodland, 1996) adaptation approach where MLLR is used to transform the features as opposed to the models (i.e. means and variances of the Gaussians). However, in comparison to that, adaptation in the proposed framework would present two distinctive advantages. First, the adaptation would by default be discriminative, i.e. learned by improving discrimination between the phone classes. Second, upon availability of more adaptation data both the feature stage and classifier stage can be adapted in a straightforward manner.
3. Sequence discriminative training: In the present work, the CNNs and MLPs were trained with frame level cross entropy criteria. It has been observed that sequence discriminative training such as maximum mutual information (MMI) or the state-level minimum Bayes risk (sMBR) criterion applied after cross entropy criteria-based training boosts ASR system performance, for example see Vesely et al. (2013). Further investigations are needed to ascertain the benefit of such sequence discriminative training applied in the proposed CNN-based framework. Along this direction we would like to also point to CRF-based end-to-end phone sequence recognition work reported in Palaz et al. (2013a) and Palaz (2016, Chapter 7), where the proposed CNN-based approach has been found to yield better system than conventional cepstral feature based approach.
4. End-to-end sequence prediction: in this article, we focused on an acoustic modeling approach where time local information $P(i|\mathbf{s}_t^c)$ is estimated in an end-to-end manner. In our recent works, we have shown that the proposed approach can be extended using conditional random fields to perform end-to-end phoneme sequence recognition (Palaz et al. (2013a) and Palaz (2016, Chapter 7)). However, performing full-fledged speech recognition through end-to-end sequence prediction is not trivial. One of the main reasons being that to search effectively and efficiently the word hypothesis the relationship between words need to be learned or modeled. As evident from the present state-of-the-art HMM-based approach, the textual data that is needed to learn the relationship between words is very different than the textual data contained in the acoustic model training data. So, joint optimization of the acoustic model and the decoder in end-to-end manner from scratch using a common data set is a highly challenging problem, and is an up-and-coming research direction (Graves and Jaitly, 2014; Amodei et al., 2015; Lu et al., 2016).
5. Going beyond conventional short-term speech signal processing: in the proposed approach one of the novelties in comparison to similar existing approaches is that short-term windowing and spectral processing mechanism is determined during training in a data- and

task-dependent manner. As a consequence of that, we found results that while challenging our understanding about short-term speech signal processing based on Fourier transform provide a link to alternate sparse coding and dictionary learning based signal processing methods. Thus, the approach opens the door to go beyond conventional short-term processing and gain further understanding about the speech signal. In that direction, at Idiap in an on-going work, the second author is involved in building over the present work to learn novel features for speaker recognition (Muckenhirn et al., 2018) and for developing countermeasures for spoofing or presentation attack detection (Muckenhirn et al., 2017).

Acknowledgments

This work was supported by the HASLER foundation (www.haslerstiftung.com) through the grant "Universal Spoken Term Detection with Deep Learning" (DeepSTD, project no. 11103 and DeepSTD-ext, project no. 14050). This work in its entirety was carried out at the Idiap Research Institute, Martigny, Switzerland, an independent not-for-profit research foundation. The authors would like to thank Dr. Ramya Rasipuram and Marzieh Razavi for their help with setting up Wall Street Journal and Mediaparl studies, and S Pavankumar Dubaganta for the help with additional experiments for rebuttal. The authors would also like to thank Dr. Michael Liebling for his critical inputs regarding the analysis of the first convolution layer in Section 5.1.3. We would like to thank the reviewers for their valuable time and feedback.

References

Abdel-Hamid, O., Deng, L., Yu, D., Jiang, H., 2013. Deep segmental neural networks for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1849–1853.

Abdel-Hamid, O., Jiang, H., 2013. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 1248–1252.

Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4277–4280.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B. C., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A. Y., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A. Y., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., Zhu, Z., 2015. Deep speech 2: end-to-end speech recognition in english and mandarin. arXiv:1512.02595.

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., et al., 2007. Greedy layer-wise training of deep networks. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), 19, p. 153.

Biem, A., Katagiri, S., McDermott, E., Juang, B.-H., 2001. An application of discriminative feature extraction to filter-bank-based speech recognition. IEEE Trans. Speech Audio Process. 9 (2), 96–110.

Bottou, L., 1991. Stochastic gradient learning in neural networks. In: Proceedings of Neuro-Nmes 91. EC2, Nimes, France.

Bourlard, H., Morgan, N., 1994. Connectionist Speech Recognition: A Hybrid Approach, 247. Springer.

Bridle, J., 1990. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In: Fogelman-Soulié, F., Hérault, J. (Eds.), Neuro-computing: Algorithms, Architectures and Applications. Springer-Verlag, New York, pp. 227–236.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS).

Collobert, R., 2004. Large Scale Machine Learning. Université de Paris VI Ph.D. thesis.

Collobert, R., Kavukcuoglu, K., Farabet, C., 2011. Torch7: a Matlab-like environment for machine learning. BigLearn, NIPS Workshop.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., 2011. Natural language processing (almost) from scratch. J. Mach. Learn. Res. 12, 2493–2537.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (1), 30–42.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Deng, L., O'Shaughnessy, D., 2003. Speech Processing: A Dynamic and Optimization-Oriented Approach. CRC Press.

Ephraim, Y., Roberts, W.J.J., 2005. Revisiting autoregressive hidden Markov modeling of speech signals. IEEE Signal Process. Lett. 12 (2), 166–169.

Furui, S., 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 11. IEEE, pp. 1991–1994.

Gales, M., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Comput. Speech Lang. 10 (4), 249–264.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., 1993. DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech disc 1-1.1. NASA STI/Recon Technical Report N 93.

Golik, P., Tüske, Z., Schlüter, R., Ney, H., 2015. Convolutional neural networks for acoustic modeling of raw time signal in LVCSR. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 26–30.

Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 1764–1772.

Graves, A., Mohamed, A., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6645–6649.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv:1512.03385.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87, 1738.

Hermansky, H., 1998. Should recognizers have ears? Speech Commun. 25 (1), 3–27.

Hifny, Y., Renals, S., 2009. Speech recognition using augmented conditional random fields. IEEE Trans. Audio Speech Lang. Process. 17 (2), 354–365.

Hillenbrand, J., Getty, L., Clark, M., Wheeler, K., 1995. Acoustic characteristics of American english vowels. J. Acoust. Soc. Am. 97 (5 part 1), 3099–3111.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. Signal Process. Mag. IEEE 29 (6), 82–97.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Hönig, F., Stemmer, G., Hacker, C., Brugnara, F., 2005. Revising perceptual linear prediction (PLP). In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2997–3000.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2 (5), 359–366.

Imseng, D., Bourlard, H., Caesar, H., Garner, P.N., Lecorv, G., Nanchen, A., others, 2012. MediaParl: bilingual mixed language accented speech database. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 263–268.

Jaitly, N., Hinton, G., 2011. Learning a better representation of speech soundwaves using restricted Boltzmann machines. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5887.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), pp. 1106–1114.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, H., Pham, P., Largman, Y., Ng, A.Y., 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Advances in Neural Information Processing Systems 22, pp. 1096–1104.

Lee, K.F., Hon, H.W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 37 (11), 1641–1648.

Lu, L., Kong, L., Dyer, C., Smith, N. A., Renals, S., 2016. Segmental recurrent neural networks for end-to-end speech recognition. arXiv:1603.00223.

Mesot, B., Barber, D., 2008. Switching linear dynamical systems for noise robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 15 (6), 1850–1858.

Mohamed, A., Dahl, G., Hinton, G., 2009. Deep belief networks for phone recognition. NIPS Workshop on Deep Learning for Speech Recognition and Related Applications.

Mohamed, A., Dahl, G., Hinton, G., 2012. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20 (1), 14–22.

Muckenhirn, H., Magimai.-Doss, M., Marcel, S., 2017. End-to-end convolutional neural network-based voice presentation attack detection. In: Proceedings of International Joint Conference on Biometrics.

Muckenhirn, H., Magimai.-Doss, M., Marcel, S., 2018. Towards directly modeling raw speech signal for speaker verification using CNNs. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 807–814.

Palaz, D., 2016. Towards End-to-End Speech Recognition. Ecole Polytechnique Fédérale de Lausanne. Thése EPFL n 7054.

Palaz, D., Collobert, R., Magimai.-Doss, M., 2013. End-to-end phoneme sequence recognition using convolutional neural networks. NIPS Deep Learning Workshop.

Palaz, D., Collobert, R., Magimai.-Doss, M., 2013. Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Palaz, D., Magimai Doss, M., Collobert, R., 2014. Learning linearly separable features for speech recognition using convolutional neural networks. arXiv:1412.7110.

Palaz, D., Magimai-Doss, M., Collobert, R., 2015. Analysis of CNN-based speech recognition system using raw speech as input. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Palaz, D., Magimai.-Doss, M., Collobert, R., 2015. Convolutionalneural networks-based continuous speech recognition using raw speech signal. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Paul, D.B., Baker, J.M., 1992. The design for the Wall Street Journal-based CSR corpus. In: Proceedings of the Workshop on Speech and Natural Language, pp. 357–362.

Poritz, A., 1982. Linear predictive hidden Markov models and the speech signal. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 7, pp. 1291–1294.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Inc..

Rabiner, L.R., Schafer, R.W., 1978. Digital Processing of Speech Signals. Prentice Hall.

Rath, S.P., Povey, D., Veselý, K., Cernocký, J., 2013. Improved feature processing for deep neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 109–113.

Razavi, M., Magimai.-Doss, M., 2014. On recognition of non-native speech using probabilistic lexical model. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Razavi, M., Rasipuram, R., Magimai.-Doss, M., 2014. Onmodeling context-dependent clustered states: comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Robinson, T., 1994. An application of recurrent nets to phone probability estimation. IEEE Trans. Neural Netw. 5, 298–305.

Sainath, T.N., Mohamed, A., Kingsbury, B., Ramabhadran, B., 2013. Deep convolutional neural networks for LVCSR. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614–8618.

Sainath, T.N., Weiss, R.J., Senior, A., Wilson, K.W., Vinyals, O., 2015. Learning the speech front-end with raw waveform CLDNNs. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH).

Schroeder, M.R., Atal, B.S., 1985. Code-excited linear prediction (CELP): high-quality speech at very low bit rates. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85, 10. IEEE, pp. 937–940.

Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 437–440.

Sheikhzadeh, H., Deng, L., 1994. Waveform-based speech recognition using hidden filter models: parameter selection and sensitivity to power normalization. IEEE Trans. Speech Audio Process. 2 (1), 80–89.

Swietojanski, P., Ghoshal, A., Renals, S., 2014. Convolutional neural networks for distant speech recognition. IEEE Signal Process. Lett. 21 (9), 1120–1124.

Swietojanski, P., Li, J., Renals, S., 2016. Learning hidden unit contributions for unsupervised acoustic model adaptation. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (8), 1450–1463. doi:10.1109/TASLP.2016.2560534.

Tüske, Z., Golik, P., Schlüter, R., Ney, H., 2014. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore, pp. 890–894.

Vesely, K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2345–2349.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K., 1989. Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. Speech Signal Process. 37 (3), 328–339.

Woodland, P., Odell, J., Valtchev, V., Young, S., 1994. Large vocabulary continuous speech recognition using HTK. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), ii. II/125–II/128 vol.2

Yegnanarayana, B., Veldhuis, R.N.J., 1998. Extraction of vocal-tract system characteristics from speech signals. IEEE Trans. Speech Audio Process. 6 (4), 313–327.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2002. The HTK Book. Cambridge University Engineering Department 3.

Yousafzai, J., Cvetkovic, Z., Sollich, P., 2009. Tuning support vector machines for robust phoneme classification with acoustic waveforms. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 2391–2394.

Zeiler, M.D., Ranzato, M., Monga, R., Mao, M.Z., Yang, K., Le, Q.V., Nguyen, P., Senior, A.W., Vanhoucke, V., Dean, J., Hinton, G.E., 2013. On rectified linear units for speech processing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3517–3521.