COMBINATION OF TWO-DIMENSIONAL COCHLEOGRAM AND SPECTROGRAM FEATURES FOR DEEP LEARNING-BASED ASR

Andros Tjandra[1,2], *Sakriani Sakti*[1], *Graham Neubig*[1], *Tomoki Toda* [1], *Mirna Adriani*[2], *Satoshi Nakamura*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]Faculty of Computer Science, Universitas Indonesia, Indonesia

andros@ui.ac.id, mirna@cs.ui.ac.id, {ssakti,neubig,tomoki,s-nakamura}@is.naist.jp

ABSTRACT

This paper explores the use of auditory features based on cochleograms; two dimensional speech features derived from gammatone filters within the convolutional neural network (CNN) framework. Furthermore, we also propose various possibilities to combine cochleogram features with log-mel filter banks or spectrogram features. In particular, we combine within low and high levels of CNN framework which we refer to as low-level and high-level feature combination. As comparison, we also construct the similar configuration with deep neural network (DNN). Performance was evaluated in the framework of hybrid neural network - hidden Markov model (NN-HMM) system on TIMIT phoneme sequence recognition task. The results reveal that cochleogram-spectrogram feature combination provides significant advantages. The best accuracy was obtained by high-level combination of two dimensional cochleogram-spectrogram features using CNN, achieved up to 8.2% relative phoneme error rate (PER) reduction from CNN single features or 19.7% relative PER reduction from DNN single features.

Index Terms— Deep learning, feature combination, cochleogram, DNN and CNN

1. INTRODUCTION

Defining acoustic features that reflect important information within utterances is one of the most critical steps in automatic speech recognition (ASR). Various feature extraction techniques have been proposed, but the acoustic features most commonly used in the conventional Gaussian mixture model - hidden Markov model (GMM-HMM) are still Mel frequency cepstral coefficients (MFCC) [1]. This is because their individual components are not strongly correlated, so it is possible to model the features using a mixture of Gaussians with diagonal covariance matrices [2].

A resurgence of deep learning has revitalized the use of the neural network paradigm for ASR. Deep neural network - HMM (DNN-HMM) hybrid systems have been proven to be superior compared to the conventional GMM-HMM model [3]. As DNNs are less sensitive to data correlation and the increase in the input dimensionality than GMMs, they allow us to exploit a richer set of features. In particular, the use of DNN with log mel-filter bank features have been shown to provide improvements in recognition accuracy [4]. Recent research has also shown that auditory features based on gammatone filters are promising to improve robustness of ASR systems [5]. Similar to MFCC, this feature is usually referred to as gammatone frequency coefficient cepstra (GFCC) [6].

Another alternative to DNNs is the use convolutional neural networks (CNNs). In speech research, CNNs were originally known as time-delay neural network [7]. By combining many convolutional and pooling layers, CNNs can be used to learn spatial-temporal patterns, while allowing robustness to translational variance in the input signals [8, 9]. CNNs with two dimensional log-mel filter banks or spectrogram input features have shown improvements over DNNs [10]. Although, CNN framework has shown to give many advantages, various features and combination within CNN framework have not been widely explored.

In this work, we attempt to explore the two dimensional features derived from gammatone filter, which are also called cochleograms within NN-HMM framework. Furthermore, we also investigated the possibilities to combine cochleogram features with spectrogram features. In particular, we combine within low and high levels of CNNs, which we call low-level and high-level feature combination. As comparison, we also construct the similar configuration with DNN in which the features were vectorized into one dimensional features.

2. BACKGROUND

2.1. Spectrogram

Spectrogram is a 2D time-frequency representation of the input speech signal. It is usually obtained via a fast Fourier transform (FFT). In this work, we use a variant of traditional spectrogram known as mel-spectrogram that commonly used in deep-learning based ASR [11, 12]. Here, the mel-scale of overlapping triangular filters are applied to the magnitude-spectrum.

2.2. Deep Neural Network

DNN is a neural network which has many hidden layers between input and output layers. Compared to traditional neural networks with one layer, DNNs have a greater capacity to learn and generalize to more complex datasets [13]. However, previously deep architectures is not intensively used for machine learning task because the difficulty for training. Several other ideas has been proposed, such as using generative pretraining like stacked denoising autoencoders (SDAE) [14], restricted Boltzmann machines (RBM) [15] to initialize the weights rather than using random initialization, or using regularization such as Dropout [16].

In this paper, we use DNNs with SDAE for generative pretraining and standard backpropagation for finetuning our models. We use a softmax output function for mapping input into each phoneme posterior probability.

2.3. Convolutional Neural Network

CNNs are neural networks that combine values between local receptive fields, shared weights, and perform sub-sampling [9]. CNN usually has one or more convolution and pooling layers. Convolutional

layers consist of multiple filters which convoluted across a given input or previous layer output. Pooling layers try to sub-sample into a lower dimension by choosing an activated unit from multiple activated units in a certain window. Using convolution and pooling, CNN has spatial-temporal connectivity and local translation invariance for the given input. In the top of several convolution and pooling layers, CNNs use fully connected hidden layers to combine the features extracted from previous convolution and pooling layers to do classification or regression.

CNN architectures have achieved high accuracy for 2D datasets that contain spatial temporal information, especially for image recognition tasks [17]. CNN architectures has advantage specially for input with strong 2D local structure (such as images or spectral representations of speech) because they are spatially and temporally highly correlated. Another advantage of CNN over fully connected neural network is CNN has weight sharing technique which can reduce the number of free parameters and improving its generalization ability [8].

3. THE USE OF COCHLEOGRAM FEATURES

3.1. Cochleogram

Cochleogram construct a time-frequency representation of the input signal to mimicking the components from the cochlea of human hearing system. To construct a cochleogram, gammatone filter is used:

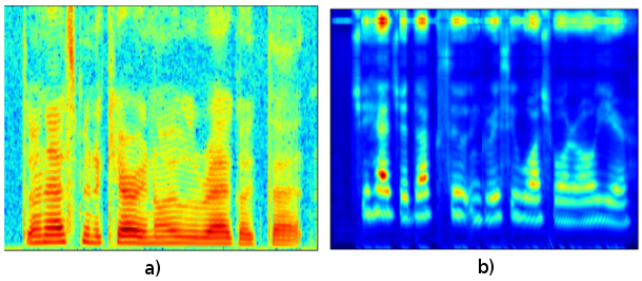
$$g(t) = \frac{at^{n-1}\cos(2\pi f_c t + \phi)}{e^{2\pi bt}} \quad (t \geq 0), \tag{1}$$

where a defines the value for amplitude, n defines the order of the filter, b defines the bandwidth, f_c is the central frequency (in kHz) and ϕ for phase (which usually we set into 0). According to [18], b is predefined:

$$b = 1.019 * 24.7 * (4.37 * f_c + 1). \tag{2}$$

In our experiments, we down-sample the frequency into frequency bands with equivalent rectangular bandwidth (ERB) scale. After we define the combination of the gammatone filterbank, we apply the filterbank into the raw speech dataset to generate a cochleogram which represents transformed raw speech in time and frequency domain.

Basically, the main difference between spectrogram and cochleogram is that cochleogram features based on ERB scale that has finer resolution at low frequencies than the mel scale used in spectrogram. Visualization comparisons of both spectrogram and cochleogram can be seen in Figure 1.



a) b)

Fig. 1. a) Spectrogram b) Cochleogram

4. FEATURE COMBINATION

4.1. Low Level Features Combination

Given an input utterance, we convert the speech into a 2D feature representation. In our case, we convert the speech into mel-filterbank spectrogram and cochleogram. We define the mel-filterbank features as matrix $x^{spec} \in \mathbb{R}^{f \times t}$ and cochleogram features as matrix $x_{coch} \in \mathbb{R}^{f \times t}$ where f represents frequency and t represents time window. In this approach, we do concatenation and the result is matrix features $x^{comb} \in \mathbb{R}^{2f \times t}$.

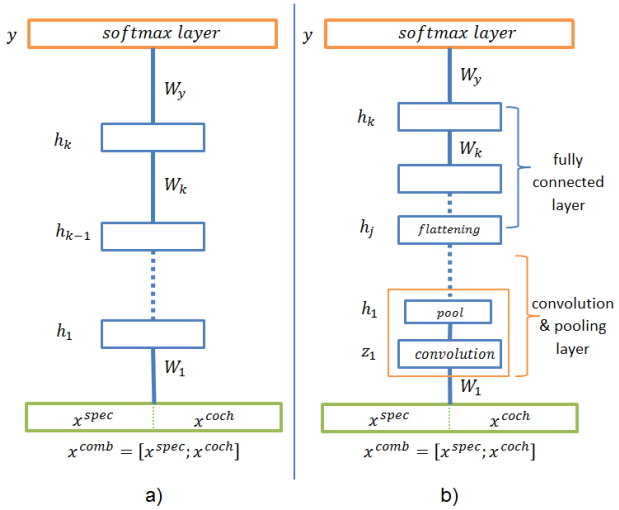


Fig. 2. a) Low level feature combination for DNN b) Low level features combination for CNN

Figure 2.a shows the detail for the DNN with low level feature combination. We vectorized the matrix features into 1D vector $\mathbb{R}^{2ft}$, then used a using Stacked Denoising Autoencoder (SDAE) to pretrain the weights $W = [W_1, ..., W_k]$ where k is the total number of hidden layers. After pretraining finished, we perform finetuning by using backpropagation to adjust the weights $[W_1, ..., W_k, W_y]$ in order to maximize the likelihood from the softmax layer. Each i-th hidden layer is calculated by using previous layer output, starting with $h_0 = x$:

$$h_i = \sigma(h_{i-1}W_i + b_i), \tag{3}$$

where W_i is a weight matrix and b_i is a bias.

Figure 2.b shows the detail for the CNN with low level feature combination. We use x^{comb} as 2D input feature then use multiple convolutional layer and max-pooling layer. Each i-th hidden layer is calculated by using previous layer output, starting with $h_0 = x$:

$$z_i = \sigma(h_{i-1} * W_i + b_i) \tag{4}$$

where W_i is a 3 dimensional vector (number of feature map, width, height) and b_i is bias for each feature map. After that, we perform max-pooling on z_i and the result is layer h_i.

After this, we flatten the last layer h_k into a vector and feed it into a fully connected hidden layer with the softmax layer. For the training step, we do not do any pretraining for any layers. The parameters $[W_1, ..., W_k, W_y]$ are trained by using the backpropagation algorithm.

4.2. High Level Features Combination

In high level feature combination, instead of using concatenation for combining input features directly, we split them into 2 different stacks of hidden layers. For DNN models in Figure 3.a, we separate input features and build two stacks of several hidden layers. On the left stacks (from h_1^{spec} to h_j^{spec}), the weight param-

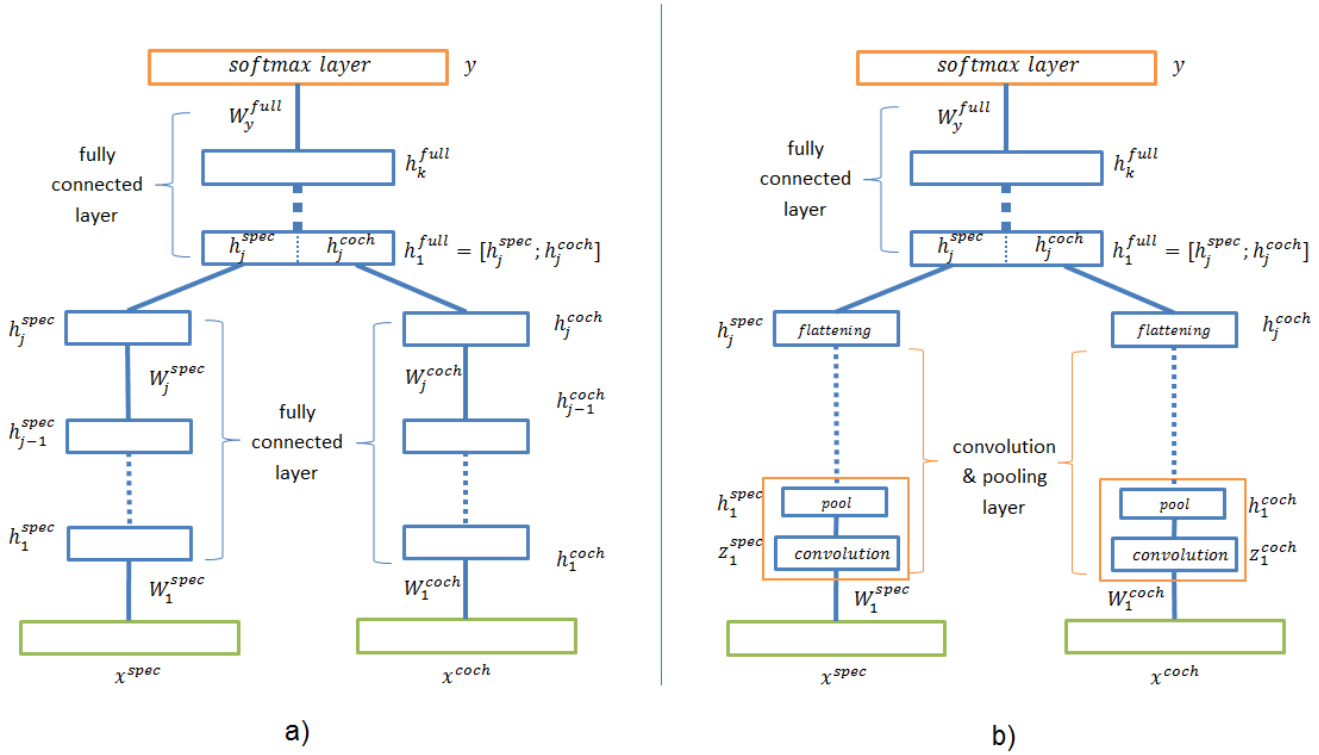


Fig. 3. a) High level feature combination for DNN b) High level features combination for CNN

eters $[W_1^{spec}, ..., W_j^{spec}]$ are pretrained by using x^{spec} input features with the Denoising Autoencoder. The same step and architecture applied for the right stacks (from h_1^{coch} to h_j^{coch}), but we change the input features into x^{coch} and pretrain the weight parameters $[W_1^{coch}, ..., W_j^{coch}]$.

After we build the pretraining model for each stack, we build the whole discriminative model by concatenating h_j^{spec} and h_j^{coch} into h_1^{full}. For the last layer, we train the softmax layer to output probabilities for each phoneme state. Finetuning was done by using backpropagation algorithm.

The same architecture is also applied in the high-level CNN model as shown in Figure 3.b. We build 2 stacks of convolutional and pooling layers. However, we do not use pre-training step for initializing the weight parameters. The training was done by using backpropagation.

5. EXPERIMENTAL SETUP

5.1. Corpus

Phone recognition experiments were perfomed on the TIMIT [1] corpus dataset. The training set contains 3696 sentences from 462 speakers. Another set of 50 speakers was used for the development set. Our model was evaluated with core data test which is consisted of 192 utterances, 8 each from 24 speakers, excluding the development set.

5.2. Front-End

We extracted the context window by using a 25-ms Hamming window with 10-ms step size. Then, the spectrogram and cochleogram speech features are generated by a Fourier-transform-based filterbanks and gammatone filter, as described in Section 2.1 and 3.1, respectively.

In our experiments, we set gammatone filter parameter into 29 frequency bands from 20 Hz to 20.000 Hz, into equivalent rectangular bandwidth (ERB) scale. For each moving window result, we average across time domain then we apply 14 context window to the left and right. This will produce a cochleogram with 29 x 29 sizes. For mel-spectrogram features, we also use 29 frequency bands from 20 Hz to 8.000 Hz with 14 context window to the left and right. Then we produce a spectogram with 29 x 29 sizes too.

5.3. Framework

Our ASR experiments were done based on the Kaldi speech recognition toolkit [19], with the DNN and CNN baseline that used a single feature stream being established based on Kaldi+PDNN toolkit [20]. For constructing DNN and CNN with low-level and high-level combination, Theano [21] libraries are used.

For DNN low-level feature combination, we use 6 fully connected hidden layer and softmax layer on the top. For DNN high-level feature combination, we use 2 different stacks of 5 fully connected hidden layer, 1 fully connected for transition between high level feature with softmax layer, and softmax layer on the top. For CNN low-level feature combination, we use 2 convolution and pooling layer and 2 fully connected hidden layer with softmax layer on the top. For CNN high-level feature combination, we use 2 different stacks of 2 convolution and pooling layer and 2 fully connected hidden layer with softmax layer on the top.

Following the TIMIT s5 recipe in Kaldi, the acoustic model consists of 1946 tied triphone states, and a phoneme-based bigram language model, estimated from the training set, is used in decoding.

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1

Then, we construct hybrid DNN-HMM and CNN-HMM systems for phoneme sequence recognition task. Here the HMM tied triphone states are used as the neural network target class. Note that, this is different than several published DNN and CNN experiments on TIMIT data, that used 183 monophone states as target classes [22, 10]. The reason is because the results from these experiments will later on be used in our triphone-based word recognition.

6. EXPERIMENT RESULTS

Table 1 shows performance comparisons of various systems in terms of phoneme error rates (PER) on TIMIT core test set. First, the DNN baseline using standard vectorized 1D mel-scale spectrogram resulting 26.58% PER, while the CNN with 2D mel-scale spectrogram perform much better achieving 23.24%. Next, applying cochleogram features on both CNN and DNN provided competitive results with CNN and DNN using mel-scale spectrogram, respectively.

After that, we performed low-level and high-level features combination as described in Section 4. As can be seen, both low-level and high-level features combination within DNN and CNN framework provided improvements in recognition accuracy. The best performances are 21.34% which was obtained by high-level combination of two dimensional cochleogram-spectrogram features within CNN framework.

Table 1. Comparisons of DNN and CNN using different features in terms of phoneme error rates on TIMIT core test set.

Model	Features	PER (%)
DNN	Mel-spectrogram	26.58
DNN	Cochleogram	26.78
DNN	Mel-Spec + Coch (Low)	26.02
DNN	Mel-Spec + Coch (High)	24.89
CNN	Mel-spectrogram	23.24
CNN	Cochleogram	23.65
CNN	Mel-Spec + Coch (Low)	22.61
CNN	Mel-Spec + Coch (High)	21.34

Overall, the combination of spectrogram and cochleogram features provided consistent improvements over single features. We hypothesize that this may be because cochleogram with ERB scale of the gammatone filter could support the better representation at lower frequency. Therefore, combining the strengths of spectrogram and cochleogram features into a single system, lead to a more accurate final result.

7. RELATED WORKS

Multiple feature streams [23] is a technique which seeks to capitalize upon practical differences between feature streams by using several features at once, in order to integrate multiple time scales in the recognition process. The basic argument is that a wide variety features have been proposed with different strengths and weakness, but the final goal is to have an ideal set of features, that reflect the important information in a consistent and well-distinguished fashion while minimizing irrelevant information. Therefore by combining the strengths of several different features into a single system, we will often obtain a more accurate final result.

Several approaches of feature combination have been devised to improve the accuracy of speech recognition systems. Classical way is to include delta and double-delta cepstral features as additional information to the static cepstral features [24]. Study by [25] explore combinations of MFCC with other features, such as perceptual linear prediction (PLP) at several levels (probability, lattice, hypothesis) within HMM-GMM ASR system, and revealed that probabilities (acoustic likelihoods) combination provide the best performance. Significant reductions in word error rate (WER) are also achieved when combining MFCC with a set of phase features [26] or voiced-unvoiced features [27] within linear discriminative analysis (LDA) based feature combination. Recently, the work by [28] presents also a multistream framework for ASR that integrates multiple streams spanning slow versus fast dynamics of speech, both spectrally and temporally.

The use of gammatone filters in combination with other features has been proposed by [29] using a bag-of-features or codebook for event detection/classification. Gammatone filter is reported to give a good approximation of the human auditory filter, and the work by [30] also showed that gammatone features lead to competitive results in large vocabulary ASR. Furthermore, different methods to combine gammatone features with a number of standard acoustic features, i.e. MFCC, PLP, Mel-Frequency PLP (MF-PLP) features, as well as MFCC-based Vocal Tract Length Normalization (VTLN) plus voicedness, were investigated and showed an improvement in performance.

Within deep learning approach, parallel use of multiple feature streams, which combine MFCC with PLP or modulation-filtered spectrogram (MSG) on hybrid/ or tandem ASR system, has also shown to provide an advantage to recognition system [31, 32]. Recent study by [33] proposed DNN and CNN combination in which linear layer and convolutional layer are combined into single linear hidden layer.

However, none of these works have explored gammatone filter in two-dimensional cochleogram features. Here, we focus on the use of cochcleogram features and its combination on various level within CNN framework. By using cochleogram features and multistream neural network model, we expects the improvement for ASR performance.

8. CONCLUSION

In this paper, we explored the use of cochleogram features in the deep-learning framework. Furthermore, we also investigated various possibilities of cochleogram and spectrogram feature combination. The results reveal that 2D features with CNN performed better than 1D features with DNN. Furthermore, cochleogram-spectrogram feature combination provided significant advantages. The best accuracy was obtained by high-level combination of two dimensional cochleogram-spectrogram features using CNN, achieved up to 8.2% relative PER reduction from CNN single features or 19.7% relative PER reduction from DNN single features. In the future, we will further investigate the use of first and second derivative of spectrogram and cochleogram, the use of various different NN structures, as well as the impact on word recognition.

9. ACKNOWLEDGEMENT

Part of this work was supported by the Commissioned Research of National Institute of Information and Communications Technology (NICT) Japan, Microsoft CORE 10 Project, and JSPS KAKENHI Grant Number 26870371.

10. REFERENCES

[1] Douglas A Reynolds and Richard C Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. on Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[2] Abdel-rahman Mohamed, Geoffrey Hinton, and Gerald Penn, "Understanding how deep belief networks perform acoustic modelling," in *ICASSP*, 2012, pp. 4273–4276.

[3] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[4] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, et al., "Recent advances in deep learning for speech research at microsoft," in *ICASSP*, 2013, pp. 8604–8608.

[5] Jun Qi, Dong Wang, Yi Jiang, and Runsheng Liu, "Auditory features based on Gammatone filters for robust speech recognition," in *IEEE ISCAS*, 2013, pp. 305–308.

[6] Xiaojia Zhao and DeLiang Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *ICASSP*, 2013, pp. 7204–7208.

[7] Alex Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin J Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, 1989.

[8] Yann LeCun and Yoshua Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, 1995.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[10] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in *ICASSP*, 2012, pp. 4277–4280.

[11] Li Deng, Ossama Abdel-hamid, and Dong Yu, "Deep convolutional neural networks using heterogeneous pooling for trading-off acoustic invariance with phonetic confusion," in *ICASSP*, 2013.

[12] Arun Narayanan and DeLiang Wang, "Joint noise adaptive training for robust automatic speech recognition," in *ICASSP*, 2013, pp. 2504–2508.

[13] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[14] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[15] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, 2012.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] Brian R Glasberg and Brian CJ Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, no. 1, pp. 103–138, 1990.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[20] Yajie Miao, "Kaldi+PDNN: Building DNN-based ASR systems with kaldi and PDNN," *CoRR*, 2014.

[21] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: a CPU and GPU math expression compiler," in *SciPy*, 2010.

[22] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[23] Yimin Zhang, Qian Diao, Shan Huang, Wei Hu, Chris D. Bartels, and Jeff Bilmes, "DBN based multi-stream models for speech.," in *ICASSP*, 2003, pp. 836–839.

[24] Sadaoki Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 1, pp. 52–59, 1986.

[25] Xiang Li, *Combination and generation of parallel feature streams for improved speech recognition*, Ph.D. thesis, Carnegie Mellon University, 2005.

[26] Ralf Schlüter and Hermann Ney, "Using phase spectrum information for improved speech recognition performance," in *ICASSP*, 2001, pp. 133–136.

[27] Andras Zolnay, Ralf Schluester, and Hermann Ney, "Robust speech recognition using a voiced-unvoiced feature," in *ICSLP*, 2001, pp. 1065–1068.

[28] Sridhar Krishna Nemala, Kailash Patil, and Mounya Elhilali, "Multistream bandpass modulation features for robust speech recognition," in *INTERSPEECH*, 2011, pp. 1277–1280.

[29] Ren Grzeszick Grzeszick Axel Plinge and Gernot A. Fink, "A bag-of-features approach to acoustic event detection," in *ICASSP*, 2014, pp. 3704–3708.

[30] Ralf Schlüter, Ilja Bezrukov, Hermann Wagner, and Hermann Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in *ICASSP*, Honolulu, USA, 2007, pp. 649–652.

[31] Daniel P.W. Ellis, "Stream combination before and/or after the acoustic model," Tech. Rep., ICSI Technical Report, USA, 2000.

[32] Hynek Hermansky, Daniel P.W. Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000, pp. 1635–1638.

[33] Hagen Soltau, George Saon, and Tara N Sainath, "Joint training of convolutional and non-convolutional neural networks," in *ICASSP*, 2014, pp. 5609–5613.