一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2016-63（2016-12）

コクレオグラムとスペクトログラムを用いた深層学習音声認識

アンドロスチャンドラ[†]　サクリアニ　サクティ[†]　ミルナアドリアーニ[††]　中村　哲[†]

† 奈良先端科学技術大学院大学
†† インドネシア大学

あらまし　本論文では、対数メル尺度やスペクトログラムにコクリオグラムを加えた特徴ベクトルを用いた、Deep Neural Network（DNN）, Convolutional Neural Network（CNN）による音声認識システムを提案する。TIMIT 音素認識タスクにおいて、スペクトログラム—コクリオグラムを用いた CNN で、スペクトログラムのみを用いた、CNN に対して 8.2%、DNN に対して 19.7%の性能向上を示した。

キーワード　深層学習：素性組み合わせ、コクリオグラム

Deep Learning-based ASR using Cochleogram and Spectrogram Features Combination

Andros TJANDRA[†], Sakriani SAKTI[†], Mirna ADRIANI[††], and Satoshi NAKAMURA[†]

† Nara Institute of Science and Technology
†† Universitas Indonesia

Abstract This paper proposes various possibilities to combine cochleogram features with log-mel filter banks or spectrogram features within the DNN and CNN framework. Performance was evaluated on TIMIT phoneme sequence recognition task. 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.

Key words Deep learning, feature combination, cochleogram, DNN and CNN

1. Introduction

Deep neural network - HMM (DNN-HMM) hybrid systems have been proven to be superior compared to the conventional GMM-HMM model. 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. Recent research has also shown that auditory features based on gammatone filters are promising to improve robustness of ASR systems [1]. Another alternative to DNNs is the use convolutional neural networks (CNNs). CNNs with two dimensional log-mel filter banks or spectrogram input features have shown improvements over DNNs [2]. 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. Neural Networks

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 [3].

Another type of neural networks is CNN. CNNs are neural networks that combine values between local receptive fields, shared weights, and perform sub-sampling. In CNN, the convolutional layers consist of multiple filters which convoluted across a given input or previous layer output and the pooling layers try to sub-sample the value from certain area. Using convolution and pooling, CNN has spatial-temporal connectivity and local translation invariance for the given input.

2.1 Cochleogram

In speech recognition, spectrogram is the widely used features obtained via fast Fourier Transform. In this work, we also explore gammatone-based filter. Cochleogram construct a time-frequency representation of the input signal to mimic the components from the cochlea of human hearing system. To construct a cochleogram, gammatone filter is used:

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

where a defines the value for amplitude, n defines the order

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.

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 [4], $b = 1.019 * 24.7 * (4.37 * f_c + 1)$. In our experiments, we down-sample the frequency into frequency bands with equivalent rectangular bandwidth (ERB) scale.

3. Feature Combination

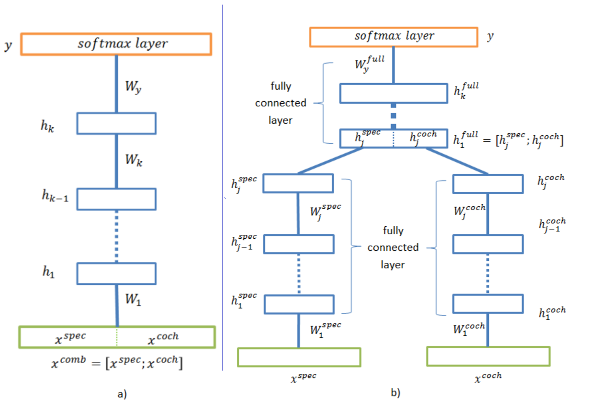

図 1 a) Low level feature combination for DNN b) High level feature combination for DNN

In low level features combination, we convert the speech into a 2D feature representation. In our case, we convert the speech into mel-filterbank spectrogram and cochleogram. In this approach, we do concatenation and the result is matrix features $x^{comb} = [x^{spec}; x^{coch}] \in \mathbb{R}^{2f \times t}$. Figure 1.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 $W = [W_1, ..., W_k]$ and followed by finetuning. To apply low-level feature combination for the CNN, we change the multiple fully connected layers into multiple convolutional and max-pooling layers respectively and feed it into a fully connected hidden layer with the softmax layer.

In high level feature combination, we split our model into 2 different stacks of hidden layers. For DNN models in Figure 1.b, we separate input features and build two stacks of several hidden layers. On the left stacks and right stacks, the weight parameters $[W_1^{spec}, .., W_j^{spec}]$ and $[W_1^{coch}, .., W_k^{coch}]$ are trained only with spectrogram and cochleogram features respectively. In the end, we concatenate h_j^{spec} and h_j^{coch} into h_1^{full} and put softmax layer. The same architecture is also applied in the high-level CNN model by replacing each fully connected layers with convolution and max-pooling layers.

4. Experimental Setup

4. 1 Corpus and Front-End

Phone recognition experiments were perfomed on the TIMIT dataset. 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 filter-banks and gammatone filter. 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. For mel-spectrogram features, we also use 29 frequency bands. Following the TIMIT s5 recipe in Kaldi, the acoustic model consists of 1943 tied triphone states.

4. 2 Framework

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.

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

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

5. Experiment Results

Table 1 shows performance comparisons of various systems in terms of phoneme error rates (PER) on TIMIT core test set. 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.

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.

6. 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. 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.

謝辞 Part of this work was supported by Microsoft CORE 10 Project as well as JSPS KAKENHI Grant Numbers 24240032 and 26870371.

文 献

[1] J. Qi, D. Wang, Y. Jiang, and R. Liu, “Auditory features based on Gammatone filters for robust speech recognition,” IEEE ISCAS, pp.305–308, 2013.

[2] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” ICASSP, pp.4277–4280, 2012.

[3] Y. Bengio, “Learning deep architectures for AI,” Foundations and trends in Machine Learning, vol.2, no.1, pp.1–127, 2009.

[4] B.R. Glasberg and B.C. Moore, “Derivation of auditory filter shapes from notched-noise data,” Hearing research, vol.47, no.1, pp.103–138, 1990.