A USEFUL FEATURE-ENGINEERING APPROACH FOR A LVCSR SYSTEM BASED ON CD-DNN-HMM ALGORITHM

Sung Joo Lee, Byung Ok Kang, Hoon Chung, and Jeon Gue Park

Speech Processing Lab., Electronics and Telecommunication Research Institute (ETRI)
Address: 161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350 South Korea
Email: {lee1862, bokang, hchung, jgp}@etri.re.kr

ABSTRACT

In this paper, we propose a useful feature-engineering approach for Context-Dependent Deep-Neural-Network Hidden-Markov-Model (CD-DNN-HMM) based Large-Vocabulary-Continuous-Speech-Recognition (LVCSR) systems. The speech recognition performance of a LVCSR system is improved from two feature-engineering perspectives. The first performance improvement is achieved by adopting the intra/inter-frame feature subsets when the Gaussian-Mixture-Model (GMM) HMMs for the HMM state-level alignment are built. And the second performance gain is then followed with the additional features augmenting the front-end of the DNN. We evaluate the effectiveness of our feature-engineering approach under a series of Korean speech recognition tasks (isolated single-syllable recognition with a medium-sized speech corpus and conversational speech recognition with a large-sized database) using the Kaldi speech recognition toolkit. The results show that the proposed feature-engineering approach outperforms the traditional Mel Frequency Cepstral Coefficient (MFCCs) GMM + Mel-frequency filter-bank output DNN method.

Index Terms— Feature extraction, feature engineering, speech recognition, deep learning, deep neural network

1. INTRODUCTION

Whereas other technologies require user effort to adapt to artificial tools, voice user interface is one of the most natural and intuitive interaction technologies for mankind. Automatic-Speech-Recognition (ASR) has been considered as a dream technology (i.e., artificial intelligence) and has been a subject in many science fiction movies for many years (e.g., HAL 9000 in *2001: A Space Odyssey*). In the world of academia, ASR has been an active research area for more than the last five decades.

However, in the past ASR was not entirely successful. Recently, the proliferation of voice-apps (e.g., voice internet search and voice messaging service) enables the massive collection of speech data. These audio footprints from ordinary people and the brilliant ideas of machine learning have fertilized the modern speech technologies. Since 2009, deep learning technology utilized by researchers has successfully replaced the Gaussian mixtures at the industrial scale [1-19]. In these days, it seems that the deep learning algorithm becomes a mainstream technology [6]. It is true that deep learning is computationally demanding. However, the recent advances in computing hardware alleviate this computational drawback.

Despite the modern technical breakthroughs in speech recognition [1-19], the recognition performance of even a state-of-the-art ASR system is still behind a human in most real-world applications. It means that puzzles in ASR are not completed even until today. Unfortunately, speech recognition still remains in a challenging research area. One of the main reasons for the difference between human and machine hearings is a capability to deal with the highly variable nature of speech. For instance, even the same speaker speaks in different styles, at different rates, and in different emotional states. The presence of environmental noise, reverberation, different microphones and recording devices results in additional variability. Therefore, the robustness of a speech recognition engine is the key to success in real-world applications.

The purpose of feature-extraction for ASR is to provide a compact vector which represents phonemic information while suppressing others (e.g., variability of speech signal which is caused by adverse environment, channel, gender, age, and so on). Therefore, the feature extractor also plays an important role in speech recognition accuracy and robustness. In this paper, we introduce a useful feature-extraction approach for CD-DNN-HMMs from two feature-engineering perspectives. In order to construct CD-DNN-HMMs, we need the HMM state-level alignment information for the supervised learning stage of the DNN [2]. However, it is almost impossible to segment huge amounts of speech data by hands. Therefore, a GMM-HMM based speech recognition system is still necessary to get the information even under the CD-DNN-HMM framework [2]. In this work, our first performance gain is caused by fine alignment information which is obtained by using the proposed feature subsets (the intra/inter-frame feature vector subsets) [20]. And the second performance gain is brought about by the proposed additional features augmenting the front-end of the DNN. Therefore, our second feature-engineering is focused on a useful feature-extraction method for the front-end of the DNN. We investigate several additional features. Among

This work was supported by the ICT R&D program of MSIP/IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning]

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

the candidate features in our study, the features based on spectral entropy, pitch information and harmonic component ratio show relatively good improvement including the probability-of-voicing (POV) in [21].

The remainder of this paper is organized as follows. In section 2, the extendable feature subsets for the state-level alignment information are expressed. And then a brief review on the additional features for the front-end augmentation of the DNN is described in section 3. In section 4, the effectiveness of the proposed features is demonstrated under the series of speech recognition tests using the popular Kaldi speech recognition toolkit [22] before the conclusion in section 5.

2. INTRA-AND INTER-FRAME FEATURE SUBSETS FOR GMM-HMM

The intra/inter-frame feature subsets are originally developed to cope with a recognition performance saturation problem in GMM-HMMs [20]. The performance saturation problem means that the recognition accuracy does not tend to be improved despite of increasing the amount of training data beyond a certain size. The reason of the saturation seems that the representation using the traditional feature set (39 dimensional MFCCs) is not sufficient. Therefore, we need more useful representations from acoustic speech data. The idea of the intra/inter-frame feature subsets is motivated by this simple idea. Figure 1 shows the schematic diagram of the feature-extraction approach proposed for GMM-HMMs.

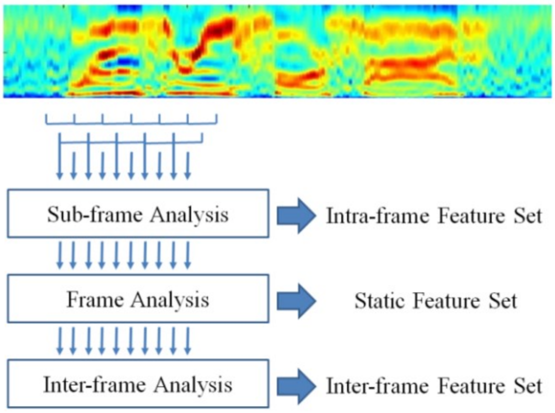

Fig. 1. The Schematic Diagram of the Proposed Feature-extraction Approach for GMM-HMMs.

As shown in figure 1, we categorize the complex characteristics of speech data in terms of signal analysis scope: static, intra-frame, and inter-frame feature subsets. The purpose of the intra-frame feature subset is to capture rapidly changing characteristics of speech spectrum. And the inter-frame feature subset represents dynamic properties of input speech along the sequential frames. Therefore, the role of the inter-frame feature subset is similar with the traditional dynamic features (i.e., delta and double-delta). The static feature subset is not different from the traditional one. It stands for average spectral envelope information in one frame. Acoustic language of a human is composed of two kinds of temporal acoustic variations. One is very rapidly changing characteristics of speech spectrum which is not fully covered by the traditional quasi-stationary assumption. And the other is temporal variations of speech along with speech-frame sequences. For instance, a vowel sound consists of specific combinations of relatively steady-state frequencies, while consonants are made of rapid transitions of frequencies that may change even within a single syllable. Not surprisingly, we (humans) are able to perceive these complex changes of speech spectrum.

In the previous work [20], we tried to extract these two temporal properties using the intra/inter-frame feature extraction framework. Rapidly changing auditory cues are captured by a simple sub-frame analysis method. More various temporal dynamics along the sequential input frames is efficiently estimated by exploiting a temporal Discrete-Cosine-Transform (DCT) method [23].

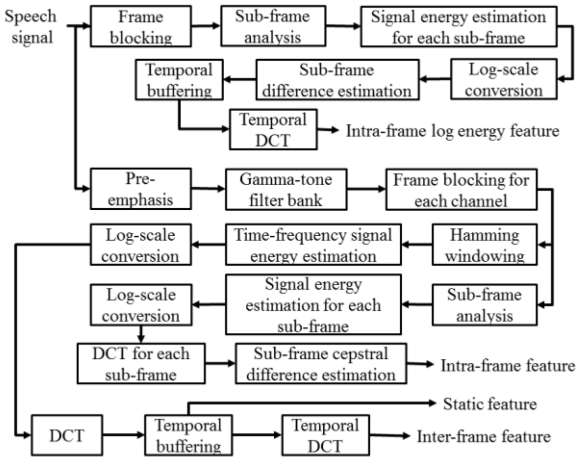

Fig. 2. The Block Diagram of the Intra/inter-frame Feature-extraction Approach.

Figure 2 shows the block diagram of the proposed intra/inter-frame feature-extraction approach. The total dimension of the proposed feature vector is extended to 69, since a sub-frame log-energy difference (1) and its temporal variations (3) are recently added to the previous 65-dimensional feature vector [20].

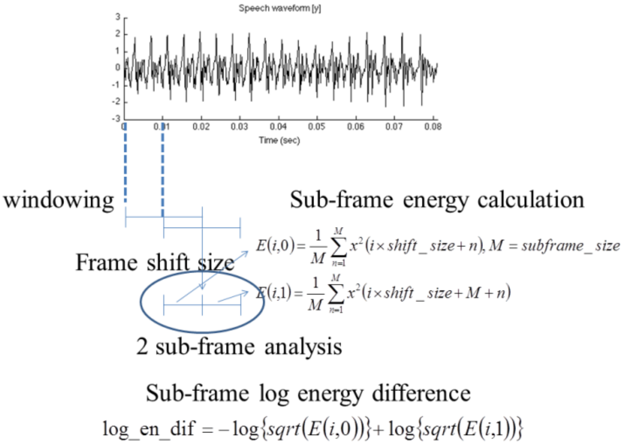
Speech waveform [y]
3
2
1
0
-1
-2
-3
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Time (sec)
windowing
Sub-frame energy calculation
$E(i,0)=\frac{1}{M}\sum_{n=1}^{M}x^2(i\times shift\_size+n), M=subframe\_size$
Frame shift size
$E(i,1)=\frac{1}{M}\sum_{n=1}^{M}x^2(i\times shift\_size+M+n)$
2 sub-frame analysis
Sub-frame log energy difference
$\log\_en\_dif=-\log\{sqrt(E(i,0))\}+\log\{sqrt(E(i,1))\}$

Fig. 3. Sub-frame Log-energy Difference Calculation.

Figure 3 indicates the estimate procedure of the sub-frame log-energy difference with the sub-frame analysis method. This feature is also useful for representing rapid changes of speech spectrum.

3. ADDITIONAL FEATURES FOR DNN INPUT

We have a question about the front-end of the traditional DNN for speech recognition. Is the Mel-frequency filter-bank output sufficient for machine hearing? In 2014, it was reported that a simple additional feature (i.e., POV) could improve the recognition performance of DNN-HMMs [21]. In this work, we investigate several additional features and introduce some useful features among them. The useful features for the front-end of the DNN are based on spectral entropy (SE), pitch information (PI) and harmonic frequency component ratio (HFCR). The spectral entropy based feature is obtained as follows. The input spectrum is converted into x_i.

$$x_i = \frac{X_i}{\max\{X_i\}} \ for \ i = 1 \ to \ N \qquad (1)$$

Where X_i indicates the energy of the $\mathrm{i_{th}}$ frequency component of the input spectrum. The entropy is computed with equation (2).

$$\mathrm{H}(x) = \sum_{i=1}^{N} x_i \times \log_{10}(x_i) \qquad (2)$$

The dynamic range of the spectral entropy values is bounded in (0, 50) and converted as follows,

$$\text{entropy feature} = \log\left(\frac{H(x)/50+0.0001}{1.0001-H(x)/50}\right) \qquad (3)$$

The proposed spectral entropy based feature is obtained with equation (3).

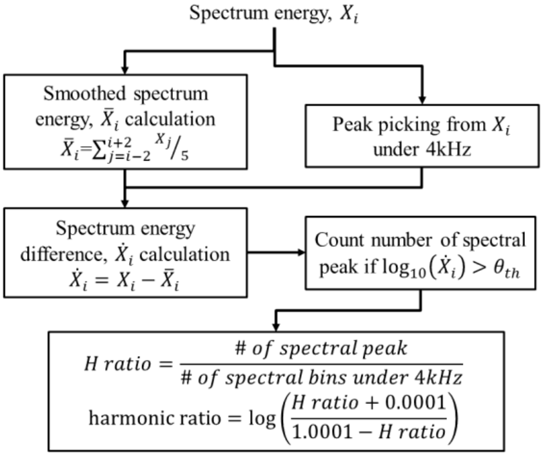
Spectrum energy, X_i
Smoothed spectrum energy, $\bar{X}_i$ calculation
$\bar{X}_i = \sum_{j=i-2}^{i+2} X_j/5$
Peak picking from X_i under 4kHz
Spectrum energy difference, $\dot{X}_i$ calculation
$\dot{X}_i = X_i - \bar{X}_i$
Count number of spectral peak if $\log_{10}(\dot{X}_i) > \theta_{th}$
$H\ ratio = \frac{\#\ of\ spectral\ peak}{\#\ of\ spectral\ bins\ under\ 4kHz}$
$\text{harmonic ratio} = \log\left(\frac{H\ ratio + 0.0001}{1.0001 - H\ ratio}\right)$

Fig. 4. Feature-Extraction Procedure of the Proposed HFCR.

The proposed HFCR represents how many harmonic frequency components are included in the input spectrum. In this work, we heuristically detect harmonic frequency components in speech spectrum as shown in figure 4. The values of the harmonic component ratio are also converted to make them good for the Kaldi DNN input.

Figure 5 describes the feature-extraction procedure of the proposed pitch information. Unlike in [21], a very simple algorithm is adopted to extract pitch information from speech input. The logarithmic value conversion approach is also applied as shown in figure 5.

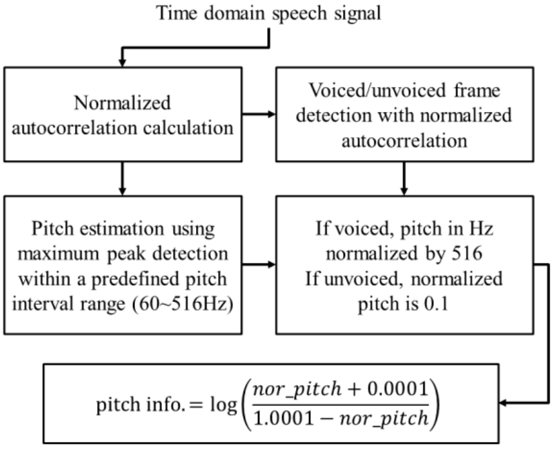
Time domain speech signal
Normalized autocorrelation calculation
Voiced/unvoiced frame detection with normalized autocorrelation
Pitch estimation using maximum peak detection within a predefined pitch interval range (60~516Hz)
If voiced, pitch in Hz normalized by 516
If unvoiced, normalized pitch is 0.1
$\text{pitch info.} = \log\left(\frac{nor\_pitch + 0.0001}{1.0001 - nor\_pitch}\right)$

Fig. 5. Feature-Extraction Procedure of the Proposed PI.

4. EXPERIMENTS

In order to demonstrate the feasibility of the proposed feature-extraction approaches, a series of speech recognition tests (isolated single-syllable recognition and conversational speech recognition tests) are conducted. We firstly evaluate the state-level alignment accuracy. And then, we investigate the effectiveness of the proposed additional features for the front-end augmentation of the DNN. Medium-sized single-syllable recognition tests are conducted to save time before the computationally demanding large-vocabulary conversational speech recognition test. The size of the training data for isolated single-syllable recognition tests is about 120 hours. And 1,713 utterances from 10 speakers (i.e. 5 males and 5 females) are prepared for the test. The test utterances are composed of isolated single-syllables including complicated Korean diphthongs. Overall Korean single-syllable recognition accuracy is relatively low, since Korean diphthongs are very hard to be identified even by a native Korean.

The feature set for the baseline speech recognition system consists of the traditional 39 dimensional MFCCs (static 13, delta 13 and double-delta 13 including C0) and 40 log-scale Mel-frequency filter-bank outputs. For the baseline GMM-HMMs, ~1,650 tied tri-phone states and ~20,060 Gaussian mixtures are obtained by using the traditional MFCCs. The input layer for DNN-HMMs has 600 (40*(7*2+1)) nodes including 7 left/right context. Linear discriminant analysis as a preprocessor is applied without dimensional reduction. The number of hidden layers is 4 and each hidden layer is composed of 1,024 nodes. The number of output layer nodes is ~1,650 depending on the tied tri-phone state number of the baseline GMM-HMMs. The exponentially decaying learning rate is adopted for 20 epochs. The initial value of the learning rate is 0.01 and the last is 0.001. Hyperbolic tangent is exploited as an activation function. Layer-wise supervised learning scheme is applied without unsupervised pre-training. We set 200k samples per each node during the training session. The dimension of the proposed feature set is extended to 69 including the intra/inter-frame feature subsets.

Feature set	Recognition rate
MFCC(GMM)→ MelFBE(DNN)	53.8%
Intra/Inter-frame (GMM)→ MelFBE(DNN)	57.6%

Table 1. Korean Syllable Recognition Result.

Table 1 shows the result of the Korean single-syllable recognition tests to compare state-level alignment accuracies. The syllable recognition result indicates that the proposed feature subsets are useful for a CD-DNN-HMM framework, since more accurate state-level alignment information is obtainable. Therefore, the absolute recognition improvement of 3.8% as shown in table 1 is achieved.

In order to further improve Korean single-syllable recognition accuracy, we propose several additional features for the front-end augmentation of the DNN.

Feature set	Recognition rate
Intra/Inter-frame (GMM)→ MelFBE+POV(DNN)	59.5%
Intra/Inter-frame (GMM)→ MelFBE+HFCR(DNN)	59.5%
Intra/Inter-frame (GMM)→ MelFBE+SE(DNN)	59.8%
Intra/Inter-frame (GMM)→ MelFBE+PI(DNN)	60.0%

Table 2. Korean Syllable Recognition Result.

Table 2 shows the result of the Korean single-syllable recognition tests to evaluate the effectiveness of the proposed additional features. In this experiment, the 69 dimensional intra/inter-frame feature vector is utilized for the state-level alignment information. The syllable recognition result shows that all the proposed additional features are useful for DNN-HMMs. As shown in table 2, the additional recognition improvement of 2.4%, the best result in table 1, is obtainable by simply adding the PI feature to the front-end of the DNN.

In order to confirm the feasibility of the proposed feature-engineering method, large-vocabulary speech recognition tests are also conducted under conversational speech recognition task. ~1,182 hour speech data (relatively clean) are prepared for the acoustic model training. All the speech data are collected by Electronics and telecommunication Research Institute in Korea. The test data are separated into 3 categories as follow,

1. SponBroadCast: interview speech in Korean broadcast news, 891 utterances
2. SponPresentation: oral presentation of Korean university students, 1,184 utterances
3. SponDebate: discussion speech in Korean broadcast discussion programs, 1,308 utterances

All the test speech samples are from ordinary persons, not a professional broadcaster or presenter. The number of hidden layers is extended to 5 with 2,048 nodes for complicated conversational speech recognition. The dictionary can cover ~489,000 unique entries.

Feature set	Set 1	Set 2	Set 3
MFCC(GMM)→ MelFBE(DNN)	75.96%	84.93%	77.61%
Intra/Inter-frame (GMM)→ MelFBE+PI(DNN)	77.90%	85.85%	79.54%

Table 3. Korean Conversational Speech Recognition Result.

Table 3 shows the Korean conversational speech recognition result between the traditional (39 MFCCs (GMM) + 40 MelFBEs (DNN)) and the proposed (69 GTCCs (GMM) + 41 MelFBEs+PI (DNN)) feature-engineering methods. The syllable recognition accuracy in table 3 indicates that the proposed feature-engineering approach is consistently effective in large-vocabulary conversational speech recognition even under the various conversional situations.

5. CONCLUSIONS

The feasibility of the proposed feature-engineering approach is demonstrated through a series of the Korean speech recognition tests (the medium-sized single-syllable recognition and the large-sized conversational speech recognition tasks) using the popular Kaldi speech recognition toolkit [22]. The recognition accuracy of acoustic models without any higher-level knowledge (e.g., language model) is evaluated under the isolated single-syllable recognition test. We think that this medium-sized syllable recognition test is an efficient assessment procedure before the time-consuming LVCSR experiment.

All the speech recognition experiment results including conversational speech recognition indicate that the proposed feature-engineering approach outperforms the traditional method. That is, the proposed intra/inter-frame feature subsets represent useful information for HMM state-level alignment. And the proposed additional features effectively augment the front-end of the DNN. In this work, all the proposed additional features for the front-end of the DNN are coincidently related with the distinct characteristics separating voicing portions from input speech. Despite of simple feature-extraction algorithms, the further recognition improvement is obtained by adding a simple feature to the front-end of the DNN. The absolute single-syllable accuracy improvement of 6.2% is totally achieved under the proposed feature-engineering approach. It is also confirmed that the effectiveness of our feature-engineering approach continues in the relatively large-vocabulary conversational speech recognition tasks.

It is reported that the intra/inter-frame feature subsets are robust in the presence of adverse noise [20]. Therefore, the further recognition performance improvement is expected in the case of matched condition training. In this experiment, any advanced techniques for GMM-HMMs are not adopted. Therefore, further performance gain is also expected by collaboration with the effective speech recognition technologies (e.g., speaker adaptation, discriminative training) [7].

REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, et al., "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Mag.*, vol. 29, no. 6, pp.82-97, Nov., 2012.

[2] G.E. Dahl, D. Yu, L. Deng, A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp.30-42, Jan., 2012.

[3] A. Graves, A.R. Mohamed and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 2013, pp.6645-6648.

[4] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp.1-13, Oct., 2014.

[5] G.E. Dahl, D. Yu, L. Deng, A. Acero, " Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 2011, pp. 4688–4691.

[6] D. Yu and Li Deng, *Automatic Speech Recognition – A Deep Learning Approach*, Springer, 2014.

[7] F. Seide, X. Chen, D. Yu, "Feature Engineering in Context-Dependent Deep Neural Networks for Conversational Speech Transcription," in *Proc. IEEE Workshop on Automatic speech Recognition and Understanding*, Honolulu, Hawaii, USA, 2011, pp.24-29.

[8] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic Modeling using Deep Belief Networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp.14-22, Jan., 2012.

[9] Y. LeCun, Y. Bengio, "Convolutional Networks for images, speech, and time series," in *Book the handbook of brain theory and neural networks*, MIT Press Cambridge, 1998, pp.255-258.

[10] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. the Neural Information Processing Systems*, Vancouver, Canada, 2006, pp. 153-160.

[11] Y. Bengio, N. Boulanger-Lewandowski, R. Pascanu, "Advances in optimizing recurrent networks," in *Proc. the International Conference on Acoustics, Speech and Signal Processing,* Vancouver, Canada, 2013, pp.8624-8628.

[12] V. Nair, and G. Hinton, "Rectified Linear Units Improve Restricted Boltzman Machines," in *Proc. The 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp.807-814.

[13] T. Mikolov, A. Deoras, D. Povey, L. Burget, J. Cernocky, "Strategies for training large scale neural network language models," in *Proc. the IEEE Workshop on Automatic Speech Recognition and Understanding*, Honolulu, Hawaii, 2011, pp. 196–201.

[14] T. Mikolov, G. Zweig, "Context dependent recurrent neural network language model," in *Proc. the IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, 2012, pp. 234–239

[15] O. Abdel-Hamid, L. Deng, and D. Yu, "Exploring Convolutional Neural Network Structures and Optimization Techniques for Speech Recognition," in *Proc. 14th Annual Conference on International Speech Communication Association*, Lyon, France, 2013, pp.3366-3370.

[16] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn, D. Yu, "Convolutional neural networks for speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 22, no.10, pp.1533-1545, Oct., 2014.

[17] H. Sak, A. Senior, F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. 15th Annual Conference on International Speech Communication Association*, Singapore, 2014, pp.338-342.

[18] D. Yu, M.L. Seltzer, "Improved bottleneck features using pre-trained deep neural networks," in *Proc. 12th Annual Conference of International Speech Communication Association*, Florence, Italy, 2011, pp. 237–240.

[19] H. Hermansky, D.P. Ellis, S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000, pp. 1635–1638.

[20] S. Lee, B. Kang, H. Chung, and Y. Lee, "Intra-and Inter-frame Features for Automatic Speech Recognition," *ETRI Journal*, vol. 36, no. 3, pp.514-517, Jun., 2014.

[21] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, S. Khudanpur, "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Florence, Italy, 2014, pp.2494-2498.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, et al., "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE Workshop on Automatic speech Recognition and Understanding*, Honolulu, Hawaii, USA, 2011.

[23] B. Milner, "A Comparison of Front-End Configurations for Robust Speech Recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Orlando, USA, 2002, vol. 1, pp. I-797-I-800.