INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Proficiency Assessment of ESL Learner's Sentence Prosody with TTS Synthesized Voice as Reference

Yujia Xiao[1,2*], *Frank K. Soong*[2]

[1]South China University of Technology, Guangzhou, China
[2]Microsoft Research Asia, Beijing, China

xiao.yujia@mail.scut.edu.cn, frankkps@microsoft.com

Abstract

We investigate how to assess the prosody quality of an ESL learner's spoken sentence against native speaker's natural recording or TTS synthesized voice. A spoken English utterance read by an ESL leaner is compared with the recording of a native speaker, or TTS voice. The corresponding F0 contours (with voicings) and breaks are compared at the mapped syllable level via a DTW. The correlations between the prosody patterns of learner and native speaker (or TTS voice) of the same sentence are computed after the speech rates and F0 distributions between speakers are equalized. Based upon collected native and non-native speakers' databases and correlation coefficients, we use Gaussian mixtures to model them as continuous distributions for training a two-class (native vs non-native) neural net classifier. We found that classification accuracy between using native speaker's and TTS reference is close, i.e., 91.2% vs 88.1%. To assess the prosody proficiency of an ESL learner with one sentence input, the prosody patterns of our high quality TTS is almost as effective as those of native speakers' recordings, which are more expensive and inconvenient to collect.

Index Terms: Nativeness, Dynamic Time Warping (DTW), Prosody, Gaussian mixture model, Deep Neural Network

1. Introduction

Learning a new language orally is always desirable when people have the business need or academic interests. While an experienced teacher plays a key role to enhance or speed up the learning process, there is usually a shortage of such teachers when the demand exceeds supply significantly. Computer Assisted Language Learning (CALL) can alleviate this problem when a well-trained computer can actively participate in the language learning process as a teacher or teaching assistant. It can evaluate a student's pronunciation at phonetic (segmental) level or prosodic (supra-segmental) level objectively to give him constructive feedbacks. Many of the Computer Aided Pronunciation Training (CAPT) systems focus on evaluating the segmental level information only. The supra-segmental information over a time span (phrase or sentence) longer than that of the segmental information (phoneme or syllable), is challenging for a beginner to produce as a native speaker.

To help a non-native language learner to learn the prosody pattern more effectively, we need to first extract prosodic features of a spoken utterance of the learner and objectively measure them against the corresponding features produced by authentic native speakers. The matching level of the prosody patterns of the native and non-native speakers can shed light on how well a learner is proficient in producing native speaker-like prosody patterns. To evaluate a speaker's nativeness at prosody level has been investigated in different ways. Based upon the mapped segmental features, Teixeira et al. [1] tried to improve the correlation between human scoring and automatic scoring by combining some global prosodic features. Sequential modeling was used in the nativeness evaluation or classification task, such as modelling over ToBI tone sequences by HMMs, bigram [2, 3] or trigram models [4]. Florian et al. proposed a large prosodic feature vector, annotation rules and prosody modelling methods (like Support Vector Regression (SVR)) [5, 6, 7]. The task of the DN Sub-Challenge of the INTERSPEECH 2015 Computational Paralinguistics Challenge is assessing the prosody of non-native English speech on a continuous scale [8]. It also provides the COMPARE feature set which contains 6,373 static features as functional of low-level descriptor (LLD) contours. In [9], Deep Rectifier Neural Network and Gaussian Processes showed a better performance than the SVM baseline. Speaker clustering has shown improved results [10].

In this paper, we use measured prosodic similarity as the feature to evaluate the nativeness of a speaker. F0 contour and breaks are the two features in characterizing the prosody of a spoken utterance. Hermes compared three methods in evaluating the similarity between two pitch contours. He found that the Fisher's Z transform of the correlation coefficient corresponded best with the auditory ratings [11]. Dynamic time warping can normalize the time duration difference between two utterances (of the same word content) [12]. In addition to the similarity between two pitch contours, in this paper we propose to measure the similarity of breaks/silences between two utterances. We found that the distributions of prosodic similarity between native speakers is distinctively different from the distribution between native and non-native speakers. The distribution of prosodic similarity measures can be well modelled by a continuous Gaussian mixture model (GMM).

Prosody annotation is a complicated and demanding work. Unlike pronunciation annotation, annotating prosody pattern requires a labeler to ignore the accuracy of the pronunciation and to focus only on the supra-segmental information, e.g. tone and rhythm. To request human experts to label prosody patterns in a consistent and accurate manner can be both time consuming and tedious [6]. Besides, even among experts, it is relatively difficult for them to agree what prosody patterns should be adopted as a golden standard. In this paper, to simplify the process, a deep neural network is trained to classify the prosody of a spoken sentence as "native" or "non-native".

We use the recordings of native speakers as reference for assessing prosody quality in our study. Additionally, we want

*Work performed as an intern in Microsoft Research Asia

http://dx.doi.org/10.21437/Interspeech.2017-64

to test if the reference utterance of native speaker can be replaced by synthesized voice of high quality Text-To-Speech (TTS) [13]. If the answer is yes, sentences can be generated on demand for any given text without going through a tedious and expensive process of collecting native speakers' utterances.

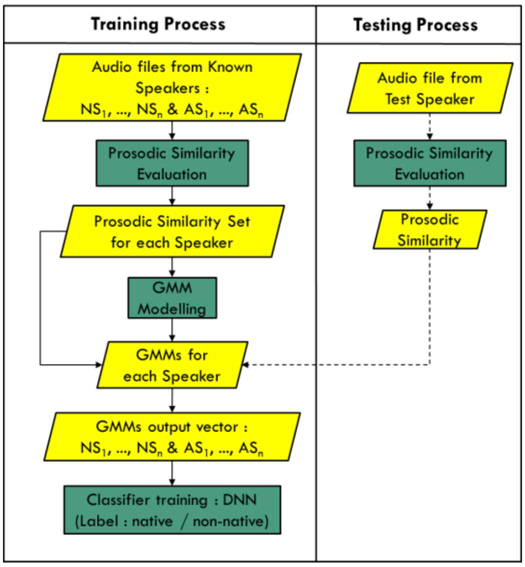
Training Process
Testing Process
Audio files from Known Speakers : NS1, ..., NSn & AS1, ..., ASn
Prosodic Similarity Evaluation
Prosodic Similarity Set for each Speaker
GMM Modelling
GMMs for each Speaker
GMMs output vector : NS1, ..., NSn & AS1, ..., ASn
Classifier training : DNN (Label : native / non-native)
Audio file from Test Speaker
Prosodic Similarity Evaluation
Prosodic Similarity

Figure 1 *Prosodic Similarity based Nativeness Evaluation, NS means native speaker and AS means non-native speaker*

Figure 1 shows the flow diagram of the modules of training and testing process used in this study. At first, we evaluate prosodic similarities of a number of utterance pairs from speakers with labels of native or non-native speakers. We then train GMMs for modelling the distribution of each speaker's prosodic similarities to native speaker(s). The GMM distributions are used to construct an input vector of a DNN classifier for classifying whether an input similarity pattern is from an utterance of native or non-native speaker.

2. Corpora

In our task, two types of utterance pair patterns are constructed. The first one is native-native utterance pair, where a sentence is spoken by two native speakers and they are used to produce the native-native prosodic similarity pattern. The second one is native-nonnative utterance pair, where the same sentence is spoken by a native speaker and a non-native speaker to produce the native-nonnative similarity pattern.

2.1. Native-native utterance pair

We use part of the data from the CMU-Arctic speech databases to construct native-native utterance pairs [14]. In this database, approximately 1,200 phonetically balanced English utterances have been carefully recorded under studio conditions by each speaker. The recordings from 2 US female native speakers (slt, clb) and 2 US male native speakers (bdl, rms) are selected in our task. We use 1,125 utterances from each speaker to construct 6,750 utterance pairs by comparing them with each other.

2.2. Native-nonnative utterance pair

Non-native speakers are the users, who have Chinese as their L1, of Microsoft English learning project "mTutor" [15]. They use it to practice speaking English, and the sentences they read after were recorded by a native speaker (a female). We will use the recording data from 4 users, 3c89 (female, 2332 utterances), a01d (female, 859 utterances), 782d (male, 1288 utterances), 9f1f (male, 1597 utterances). Altogether, we have 6,076 native-nonnative utterance pairs.

3. Prosodic Similarity Evaluation

We evaluate prosodic similarity from in both F0 and Break patterns.

3.1. Syllable-level DTW-based F0 similarity measure

This method is similar to the work in [12] but with some differences. The framework is shown in Figure 2. We extracted MFCC features and used them for forced aligning an utterance with the corresponding word sequence. With the segmentations, a syllable-level dynamic time warping was performed. The MFCC (a multi-dimensional feature) instead of F0 (a scalar feature) was used for a more reliable warping result. We then extracted F0 sequences by following the warped MFCC features. Different speakers usually have different F0 distribution. To make extracted F0 contours of different speakers comparable, F0 sequences were normalized by subtracting the average F0 on an utterance level. Finally, the correlation coefficient between two equalized F0 sequences of utterances 1 and 2 was computed for its similarity [11] in Eq.1, where N is the number of voiced frames, f_1^n is the F0 value in the n^{th} voiced frame in utterance 1, $\bar{f_1}$ is the average F0 of all voiced frames in utterance 1.

$$r_{f_1 f_2} = \frac{\sum_{n=1}^{N}(f_1^n - \bar{f_1}) * (f_2^n - \bar{f_2})}{\sqrt{\sum_{n=1}^{N}(f_1^n - \bar{f_1})^2} * \sqrt{\sum_{n=1}^{N}(f_2^n - \bar{f_2})^2}} \quad (1)$$

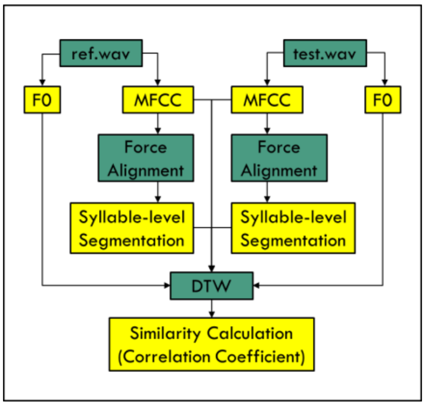
ref.wav
test.wav
F0
MFCC
MFCC
F0
Force Alignment
Force Alignment
Syllable-level Segmentation
Syllable-level Segmentation
DTW
Similarity Calculation (Correlation Coefficient)

Figure 2 *Framework of Syllable-level DTW-based F0 Similarity Measure*

3.2. Alignment-based break similarity measure

Forced alignment provides the position and duration of break(s) in an utterance. We propose a method to calculate the break similarity of two utterances. In this method, the break similarity is defined as the product of break position similarity and break duration similarity.

3.2.1. Break position matching

Utterances of the same text are compared according to the corresponding syllable sequences. For example, the sentence read by speakers A and B is "You are outgoing", we segment them into corresponding syllable sequences as "y.uw aa.r aw.t g.ow ih.ng". Therefore, 4 bi-syllabic pairs are thus constructed, which are “y.uw-aa.r”, “aa.r-aw.t”, “aw.t-g.ow”, “g.ow-ih.ng”. A break can be inserted in any such syllable pair. For each syllable pair, e.g., when there is no silence inside the “y.uw-aa.r” syllable pair, it indicates that there is no break in between; otherwise, when there is a break (silence) between them, we use “y.uw-sil-aa.r” to mark it. If the marking from speaker A is the same or different from the corresponding from speaker B, the break position for this syllable pair is counted as matched (with a value 1) or mismatched (with a value 0), correspondingly.

Four possible types of syllable pairs and the corresponding values are listed in Table 1. The percentage of the matched syllable positions out of the total number of syllable pairs is used as break position similarity. In our sentence example, if the syllable sequence is “y.uw sil aa.r aw.t g.ow sil ih.ng” from speaker A and “y.uw aa.r sil aw.t g.ow sil ih.ng” from speaker B, the break position similarity is 0.5 between A and B for this sentence.

Table 1: *Four types of compared adjacent syllable pairs and their corresponding value.*

Compared adjacent syllable pair	**Value**
syll1-sil-syll2 / syll1-sil-syll2	1
syll1-syll2 / syll1-syll2	1
syll1-sil-syll2 / syll1-syll2	0
syll1-syll2 / syll1-sil-syll2	0

3.2.2. Break duration matching

Break duration comparison is based upon the analysis of the break position. When the same syllable pair in two utterances have a silence break inside, we will also compare their break durations.

Since different speakers can have different speech rates, we need to normalize the durations of different speakers with the corresponding speech rates as in Eq. 2, where s_i is the speech rate of speaker i, d_m is the duration of the m^{th} syllable in the utterance, and M is the total number of syllables in this utterance.

$$s_i = \frac{M}{\sum_{m=1}^{M} d_m} \tag{2}$$

The break duration similarity of the k^{th} syllable pair is computed as shown in Eq. 3

$$r^k_{b_1 b_2} = \frac{s_1 * d_1^k}{s_2 * d_2^k} \tag{3}$$

where s_1 and s_2 are the speech rates of speaker 1 and 2, respectively; d_1^k and d_2^k , the break durations of the k^{th} syllable pair of speaker 1 and speaker 2. It should be noted that the smaller value is used as the numerator in Eq. 3 to constrain the value within 0 and 1. The utterance-level break duration similarity $r_{b_1 b_2}$ is the average of syllable-level break duration similarity in Eq. 4, where K is the total number of matched syllable pairs with breaks inside.

$$r_{b_1 b_2} = \frac{\sum_{k=1}^{K} r^k_{b_1 b_2}}{K} \tag{4}$$

3.3. Distribution of prosodic similarity

We extract the two prosodic similarities (F0 and break) and analyze their distributions in the two databases, CMU-Arctic and mTutor-Users. Figs 3 and 4 show the distributions of F0 similarity and break similarity histograms, respectively. In Fig. 3, we observe that F0 similarity in CMU-Arctic, i.e., native speakers, is with a higher mean than that in mTutor-User, i.e., non-native. The shape of the distributions are similar to a Gaussian one. The two corresponding distributions of break similarity are more separated from each other as shown in Fig 4. The distributions show that both F0 and Breaks prosodic features can distinguish native from non-native speaker, based upon their similarity pattern in a single utterance.

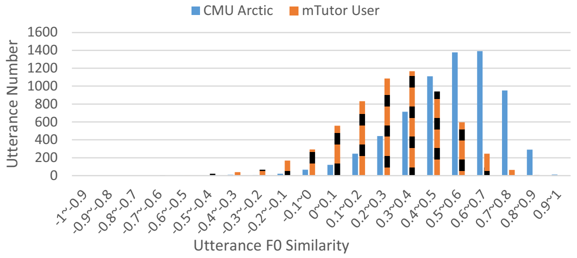

Figure 3 *The distribution of F0 similarity in datasets CMU-Arctic and mTutor-User*

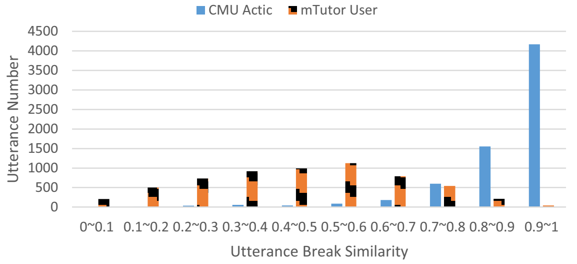

Figure 4 *The distribution of Break similarity in datasets CMU-Arctic and mTutor-User*

4. Gaussian Mixture Model

We use GMMs [16] to model the distribution of a speaker’s prosodic similarity:

$$\mathrm{p}(x|\lambda) = \sum_{c=1}^{C} w_c g(x|\mu_c, \Sigma_c) \tag{5}$$

Eq. 5 is a weighted sum of C component Gaussian densities used for the input feature in our DNN training. In the equation, x refers to the input data; λ, the model’s parameters; w_c , the weight of the c^{th} component; g , the Gaussian component, which is defined in Eq 6:

$$g(x|\mu_c, \Sigma_c) = \frac{\exp\{-\frac{1}{2}(x-\mu_c)' \Sigma_i^{-1} (x-\mu_c)\}}{(2\pi)^{D/2} |\Sigma_c|^{1/2}} \tag{6}$$

In our task, x is a 1-dimensional prosodic similarity. The number of Gaussian components is determined by Akaike information criterion (AIC) [17]. AIC provides a measure of model quality for a given set of data, given in Eq. 7, where NlogL is the negative log-likelihood of the model and q is the number of estimated parameters. The model with minimum AIC value is selected.

$$AIC = 2 * NlogL + 2 * q \tag{7}$$

5. Deep Neural Network

Deep Neural Networks (DNN) have pushed forward the speech technology in speech recognition, TTS and other speech processing [18, 19]. In this paper, a feedforward network was trained to perform a classification task. The aim is to classify a speaker as a native speaker or a non-native speaker when prosodic similarities between his spoken sentence and that of a native speaker is measured and input to the classifier.

5.1. Feedforward network

In our networks, sigmoid function is used as the activation function. We use a softmax function to convert the output to posterior probabilities. Stochastic gradient descent (SGD) [20] is used as the optimization approach to minimize the loss function (cross-entropy). Table 2 lists all DNN models we have trained in this paper. Each of them has one input layer, one hidden layer (32 hidden units) and one output layer. We set a sample-level learning rate at 0.003 for the first 50 batches, 0.002 for the next 50 batches, and 0.0001 for the rest batches. The number of epochs is 20. We used a mini batch of 50 samples in model D and E while 80 samples in other models.

5.2. Construction of the input vector

We trained GMMs for each speaker to model the distribution of a speaker's prosodic similarities. Given an utterance pair, prosodic similarity is calculated presented in Section 3. The data is used to train a speaker's GMM, the output probability density will be one component of the input vector of DNN model. Therefore, the dimension of the input vector is determined by the number of GMMs we used. Six models with different input vectors are listed in Table 2. Models A, B and C used GMMs from 4 native speakers and 4 non-native speakers. Model D used GMMs from 2 TTS (each trained with an individual speaker's recordings) and 2 non-native speakers, while model E used GMMs from 2 native speakers and the same non-native speakers used in model D. The evaluation of these models are discussed in Section 6.

6. Results

From models A to C, we selected prosodic similarities from 6,000 native-native utterance pairs and 6,000 native-nonnative utterance pairs. For model D, we selected that from 4,500 TTS-native utterance pairs and 4,500 TTS-nonnative utterance pairs. Model E is for comparison with model D, trained with prosodic similarities from 4,500 native-native utterance pairs and 4,500 native-nonnative utterance pairs. All native speakers' data are from CMU-Arctic corpus, non-native speakers' data are from mTutor-User corpus and synthesized speech is from two high quality TTS voice fonts of Microsoft TTS. Each model's datasets were randomly divided into 6 subgroups with the same size, where 5 groups were used for training a neural net-based classifier and the remaining group was used for testing. A cross-validation was performed on the 6 subgroups and the average classification accuracy was used as the final results depicted in Table 2.

6.1. F0 similarity and Break similarity

With Model A, F0 similarities are converted 8 F0-GMMs of the 8 speakers (4 native and 4 non-native). Model B is similar to Model A except it uses Br-GMMs and Break similarity data. Figs 3 and 4 have shown that break similarity performs better than F0 similarity to distinguish native from nonnative speakers. A similar trend can also be observed in results shown in Table 2. Model B (73.9%) performs better than model A (68.8%). By augmenting the two prosodic similarities together as features in model C, we improved the classification accuracy to 76.7%.

6.2. Log transformation

The input vector constructed by the output of the GMMs is a set of densities. To avoid a possible underflow in taking log, we constrain the value to a small positive floor. In Table 2, all models obtained improved performance after the log transformation. The best result is from model C, at a high classification accuracy of 91.7%.

6.3. TTS-based synthesized voice as reference

The classification performance of model D is at 88.1%, slightly lower than that of model E at 91.2% but still quite good. A high correlation coefficient of 0.957 between the posterior predicted by models D and E, computed as in Eq. 8, justifies the usage of TTS voice to replace native speakers' recordings. p_D^u is the posterior of the u^{th} utterance is spoken by a native speaker predicted by model D

$$r_{p_D p_E} = \frac{\sum_{u=1}^{U} p_D^u * p_E^u}{\sqrt{\sum_{u=1}^{U} {p_D^u}^2} * \sqrt{\sum_{u=1}^{U} {p_E^u}^2}} \tag{8}$$

Table 2: *Classification Accuracy with Log transformation*

Model No.	**Prosodic Feature**	**Input Dimension**	**Classification Accuracy**	
			RawInput (%)	+ Log (%)
A	F0	8	68.8	73.2
B	Br	8	73.9	91.0
C	F0+Br	16	76.7	91.7
D (TTS)	F0+Br	8	71.1	88.1
E	F0+Br	8	72.9	91.2

7. Conclusion

Prosodic similarities in F0 and Break are studied in this paper. They are used for classifying native English speakers from non-native ESL learners and for assessing the non-nativeness of an ESL learner. Based upon the distributions of native and non-native speakers' prosodic similarity patterns, we train deep neural nets to classify a sentence as uttered by a native or a non-native English speaker. The best classification accuracy is obtained at 91.7% by using CMU-Arctic and mTutor-User speech databases. By replacing native speakers' references with Microsoft TTS, we obtain a classification performance of 88.1%, which is fairly close to that obtained by using native speakers' recordings. The result achieved with our TTS voice is high enough to justify its usage for assessing the prosody quality of a learner's utterance spoken after a prompted text or corresponding TTS synthesized voice.

8. Acknowledgements

We would like to acknowledge our colleagues at Microsoft TTS: Xi Wang, for providing TTS technical support and many constructive comments; Fenglong Xie and Wenping Hu, for their stimulating discussions.

9. References

[1] C. Teixeira, H. Franco, E. Shriberg, K. Sonmez, K. Precoda, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," in *INTERSPEECH*, Beijing, 2000, PP. 187-190.

[2] J. Tepperman, A. Kazemzadeh, S. Narayanan, "A text-free approach to assessing nonnative intonation," in *INTERSPEECH*, Antwerp, 2007, PP. 2169-2172.

[3] A. Rosenberg, *Automatic detection and classification of prosodic events,* Columbia University, 2009.

[4] A. Rosenberg, "Symbolic and Direct Sequential Modeling of Prosody for Classification of Speaking-Style and Nativeness," in *INTERSPEECH*, 2011, PP. 1065-1068.

[5] Hönig F, Batliner A, Weilhammer K, et al, "Automatic assessment of non-native prosody for english as l2," in *Speech Prosody*, 2010, Vol. 100973, No. 1, pp. 1-4.

[6] Hönig F; Batliner A; Nöth E, "Automatic assessment of non-native prosody annotation, modelling and evaluation," *Proceedings of ISADEPT,* 2012.

[7] Hönig F, Bocklet T, Riedhammer K, et al, "The Automatic Assessment of Non-native Prosody: Combining Classical Prosodic Analysis with Acoustic Modelling," in *INTERSPEECH*, 2012, PP. 823-826.

[8] Schuller B W, Steidl S, Batliner A, et al, "The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition," in *INTERSPEECH*, 2015, PP. 478-482.

[9] Grósz T, Busa-Fekete R, Gosztolya G, et al, "Assessing the degree of nativeness and Parkinson's condition using Gaussian processes and deep rectifier neural networks," in *INTERSPEECH*, 2015, PP. 919-923.

[10] Black M P, Bone D, Skordilis Z I, et al, "Automated evaluation of non-native English pronunciation quality: combining knowledge- and data-driven features at multiple time scales," in *INTERSPEECH*, 2015, PP. 493-497.

[11] Hermes D J, "Measuring the perceptual similarity of pitch contours," *Journal of Speech, Language, and Hearing Research,* vol. 41, no. 1, pp. 73-82, 1998.

[12] Rilliard A, Allauzen A, de Mareüil P B, "Using Dynamic Time Warping to Compute Prosodic Similarity Measures," in *INTERSPEECH*, 2011, PP. 2021-2024.

[13] Yan Z J, Qian Y, Soong F K, "Rich-context unit selection (RUS) approach to high quality TTS," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. IEEE*, 2010,PP. 4798-4801.

[14] Kominek J, Black A W, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[15] http://www.engkoo.com/

[16] Reynolds D, "Gaussian mixture models," *Encyclopedia of biometrics,* pp. 827-832, 2015.

[17] Bozdogan H, "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrika,* vol. 52, no. 3, pp. 345-370, 1987.

[18] Yu D, Deng L, Automatic speech recognition: A deep learning approach, Springer, 2014.

[19] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "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.

[20] Bottou L, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010. Physica-Verlag HD*, 2010, PP. 177-186.