Vietnam J Comput Sci (2016) 3:247–257
DOI 10.1007/s40595-016-0071-3

REGULAR PAPER

Speech classification using SIFT features on spectrogram images

Quang Trung Nguyen[1] · **The Duy Bui**[1]

Received: 13 December 2015 / Accepted: 27 May 2016 / Published online: 16 June 2016

Abstract Classification of speech is one of the most vital problems in speech processing. Although there have been many studies on the classification of speech, the results are still limited. Firstly, most of the speech classification approaches requiring input data have the same dimension. Secondly, all traditional methods must be trained before classifying speech signal and must be retrained when having more training data or new class. In this paper, we propose an approach for speech classification using Scale-invariant Feature Transform (SIFT) features on spectrogram images of speech signal combination with Local naïve Bayes nearest neighbor. The proposed approach allows using feature vectors to have different sizes. With this approach, the achieved classification results are satisfactory. They are 73, 96, 95, 97 %, and 97 % on the ISOLET, English Isolated Digits, Vietnamese Places, Vietnamese Digits, JVPD databases, respectively. Especially, in a subset of the TMW database, the accuracy is 100 %. In addition, in our proposed approach, non-retraining is needed for additional training data after the training phase. The experiment shows that the more features are added to the model, the more is the accuracy in performance.

Keywords LNBNN · SIFT · Speech perception · Speech classification

✉ Quang Trung Nguyen
trungnq@vya.edu.vn

The Duy Bui
theduybui@gmail.com

[1] Human Machine Interaction Laboratory, University of Engineering and Technology, VNU Hanoi, Hanoi, Vietnam

1 Introduction

Studies on speech processing have been carried out for more than 50 years. Despite the fact that a great deal about how the system works has been researched, there is still more to be discovered. Previously, researches considered speech perception and speech recognition as separate domains. Speech perception focuses on the process that operates to decode speech sounds no matter what words those sounds might comprise.

However, there have been some differences between speech perception, speech classification and speech recognition. The differences are that speech recognition points out what the input signal is, while speech perception results in an interaction of input signal and speech classification, organizing speech signal into a category for its most effective and efficient use base on a set of training speech signal. In this paper, we focus on the problem of speech classification or, more particularly, on isolated words classification.

Researches on the speech classification have started with [1–3]. Some popular theory for speech classification are Motor theory [2], TRACE model [4,5], Cohort model [6] and Fuzzy-logical model [4].

The Motor theory was proposed by Liberman and Cooper [2] in the 1950s. The Motor theory was developed further by Liberman et al. [1,2]. In this theory, listeners were said to interpret speech sounds in terms of the motoric gestures they would use to make those same sounds.

The TRACE model [5] is a connectionist network with an input layer and three processing layers: pseudo-spectra, phoneme and word. There are three types of connection in TRACE model. The first connection type is feedforward excitatory connections from input to features, features to phonemes and phonemes to words. The second connection type is lateral inhibitory connections at the feature, phoneme

and word layers. The last connection type is top-down feedback excitatory connections from words to phonemes.

The original Cohort model was proposed in 1984 by Wilson et al. [6]. The core idea at the heart of the Cohort model is that human speech comprehension is achieved by processing incoming speech continuously as it is heard. At all times, the system computes the best interpretation of currently available input combining information in the speech signal with prior semantic and syntactic context.

The fuzzy logical theory of speech perception was developed by Massaro [4]. He proposes that people remember speech sounds in a probabilistic, or graded, way. It suggests that people remember descriptions of the perceptual units of language, called prototypes. Within each prototype, various features may combine. However, features are not just binary, there is a fuzzy value corresponding to how likely it is that a sound belongs to a particular speech category. Thus, when perceiving a speech signal our decision about what we actually hear is based on the relative goodness of the match between the stimulus information and values of particular prototypes. The final decision is based on multiple features or sources of information, even visual information.

For the speech recognition problem, some common methods are hidden Markov models (HMM) [7,8], neural network [9,10], dynamic time wrapping [11], deep neural network (DNN) acoustic models [12,13]. These approaches usually use frequent features of speech signal such as MFCC [9], LPC [14] or raw speech signal using a convolution neural network to learn features [16–18] as input features. To be used with common machine learning techniques, the size of these input features must be the same. Thus, the speech features must be resampled or quantized to have the same size. In addition, the disadvantage of these machine learning techniques is that they do not allow adding training samples without retraining. This reduces the flexibility needed for large-scale speech perception application. To retain all the discriminative features of the data, Boiman proposed a classification approach called naïve Bayes Nearest neighbor (NBNN) [19], then Sancho developed this method and proposed an approach called local naïve Bayes nearest neighbor (LNBNN) [20]. These approaches were successful in images classification problem.

In this study, we propose an approach for speech classification based on spectrogram images. In this approach, we proposed the use of scale-invariant feature transform (SIFT) of speech signal spectrogram image. SIFT features are invariant to scale and have been used well for image classification [21,22]. In particular, feature points with SIFT description are extracted successfully from 2D image of frequency spectral of speech signal. The quantity of feature points of each image is different. Each feature point describes one local feature of image; therefore, quantization of these features will result in the loss of the descriptive nature about the local feature of the image. Therefore, we should use an algorithm classification which allows using feature vectors in different sizes, while LNBNN [20] accepts input features with different sizes and has acceptable running time. In this paper, we propose the use of LNBNN for classifying speech signal represented by image based on SIFT features. This is motivated from the work in [20], where LNBNN is used with SIFT features for searching images in a database. One advantage of this approach is that new training samples are also added without retraining. Moreover, this approach allows using feature vectors in different sizes.

The paper is structured as follows. Section 2 presents related works. The scale-invariant feature transform and the local naïve Bayes nearest neighbor method is described in Sect. 3. Then, our approach is explained in Sect. 4, and experiments to show the performance of the approach are shown in Sect. 5.

2 Related works

Speech classification by machine came into existence in the early 1920s. The first machine to recognize speech was manufactured in 1920s [23]. The earliest attempts to devise systems for automatic speech classification by machine were made in 1950s, when various researchers tried to exploit the fundamental ideas of acoustic phonetics. During 1950s, most of the speech classification systems investigated spectral resonances during the vowel region of each utterance which were extracted from output signals of an analog filter bank and logic circuits [24]. In 1952, Davis et al. [25] built a system for isolated digit classification for a single speaker. The system relied heavily on measuring spectral resonances during the vowel region of each digit. In 1956, Olson et al. [26] tried to classify ten syllables of a single talker, as embodied in ten monosyllabic words. The system again relied on spectral measurements primarily during vowel regions. In 1959, Fry [27] tried to build a phoneme recognizer to recognize four vowels. They used a spectrum analyzer and a pattern matcher to make the recognition decision.

In the 1970s, research on speech classification just focused on isolated word recognition [28] and the problems of connected word recognition was a focus of research in the 1980s with the goal of creating a robust system capable of recognizing a fluently spoken string of words based on matching a concatenated pattern of individual words [29].

In the 1990s, Loizou and Cole [30,31] proposed a high-performance alphabet recognition based on context-dependent phoneme HMM. They used E-set letters consisting of the letters B, C, D, E, G, P, T, V and Z to perform the recognition experiment. In the experiment, they also addressed confusion caused by nasals (letters M and N). They have achieved a 95 % recognition rate for speaker-

independent E-set recognition and an overall alphabet recognition rate of 97.3 %. Cole and Fanty [32,33] proposed the English Alphabet Recognizer (EAR) system that performed recognition of isolated alphabets. In this system, a rule-based segmenter was used to segment the alphabet into four broad phonetic categories. Then, features were extracted from these broad phonetic categories. These features were the input of back propagation neural network (BPNN) for classification. Cole and Fanty used BPNN with conjugate gradient optimization consisting of 617 input nodes, 52 hidden layer and 26 output nodes as the classification method. The EAR system result achieved 96 % for speaker-independent letter recognition. Favero [15] also proposed speech-recognized model on E-set letters. In this model, the speech signal is parameterized with compound wavelets and an HMM-based recognizer was used for classification. Experiments were conducted by varying the compound level to note the increase in the recognition rate. The best recognition rate obtained was 71.4 % at compound level 4. However, they did not solve the problem in a noisy environment.

In 2001, Karnjanadecha [34] proposed signal modeling for high performance and robust isolated word recognition. In this model, HMM was used for classification. The recognition accuracy rate of this experiment was 97.9 % for speaker-independent isolated alphabet recognition. When adding Gaussian noise (15 dB) or testing like telephone speech simulation, the recognition rates were 95.8 and 89.6 %, respectively.

In 2004, Ibrahim [35] presented a technique to overcome the confusion problem by means of time-extended features. He expanded the duration of the consonants to gain a high characteristic difference between confusable pairs in the E-set letters. A continuous density HMM model was used as the classifier. The best recognition rate was only 88.72 %. Moreover, the author did not test on any noisy speech.

In 2011, Jonathan developed a model for Sound event classification in mismatched conditions [36]. In this model, they developed a nonlinear feature extraction method which first maps the spectrogram into a higher dimensional space, by quantizing the dynamic range into different regions, and then extracts the central moments of the partitioned monochrome intensity distributions as the feature of sound.

In 2009, Mohamed et al. tried using pre-trained, deep neural networks as part of a hybrid monophone DNN–HMM model on TIMIT, a small-scale speech task [37], and in 2012, Mohamed et al. were the first to succeed in pre-trained DNN–HMMs on acoustic modeling with varying depths of networks [38,39]. In 2013, Bocchieri and Tuske succeeded in using DNN for speech recognition for large vocabulary speech tasks [40,41].

Ossama et al. proposed using convolutional neural networks (CNN) for speech recognition in 2014. They showed that the hybrid deep neural network hidden Markov model (DNN–HMM) has been shown to significantly improve speech recognition performance over the conventional Gaussian mixture model–hidden Markov model (GMM–HMM). Their experimental results show that CNN reduces the error rate by 6–10 % compared with DNN on the TIMIT phone recognition and the voice search large vocabulary speech recognition problem.

In 2015, Palaz at al. used CNN for continuous speech recognition using raw speech signal [42]. They extended the CNN-based approach to large vocabulary speech recognition problem and compared the CNN-based approach against the conventional ANN-based approach on Wall Street Journal corpus. They also showed that the CNN-based method achieves better performance in comparison with the conventional ANN-based method as many parameters and features learned from raw speech by the CNN-based approach could generalize across different databases.

In 1997, early classification approach based on naïve Bayes classifier was proposed by Domingos and Pazzani [43] in images classification where the independence assumption was violated. They showed that the naïve Bayes classifier can perform well with regard to misclassification rate even when the independence assumption does not hold. They have performed extensive evaluations on many real-world datasets that violate the independence assumption and have showed that the classification performance is equal to or better than other learning methods. In 2008, Boiman et al. [19] proposed a feature-wise nearest neighbor model called naive Bayes nearest neighbor (NBNN). They did not quantize the descriptors, but retained all of the reference descriptors in their original form. In 2010, they optimized NBNN by correcting it for the case of unbalanced training sets [44]. They also pointed out that a major practical limitation of NBNN is the performing time in the nearest neighbor search. In 2011, Tuytelaars et al. [45] used the NBNN response vector of a query image as the input feature for a kernel SVM. This allowed discriminative training and combination with other complementary features using multiple kernels. Kernel NBNN produces increase in classification accuracy over the use of the basic NBNN algorithm.

In the NBNN model, a large amount of time needed to search for the nearest neighbors is a big problem. Even approximate methods can be slow here. They scale linearly with the number of categories. To overcome all the problems, Sancho et al. [20] introduced a local nearest neighbor improving the original NBNN in 2012. The authors figured out that the high performing time was caused by the manifold structure of descriptor space. This led to poor estimation of Euclidean distances. Their methods took advantage of local coding [46] and early cutoff soft assignment to use only the local neighborhood of a descriptor during the coding step

[47]. By restricting the coding to use only the local dictionary elements, the method has achieved improvements over their non-local equivalents.

In the above-mentioned methods, an important step is extracting certain important information from the speech signal. Feature extraction could be seen as extracting certain mathematically parameterized information from the original source signal. There are many traditional feature extraction techniques that may be used such as fast Fourier transform (FFT) coefficients [48], perceptual linear prediction (PLP) [14], linear predictive cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCCs) [7,9]. Some recent researches already have extracted speech feature from spectrogram image of a speech signal for audio retrieval system [49,50].

3 Background

In this paper, we propose using LNBNN in combination with SIFT feature of spectrogram images which are converted from speech signals. This section provides an overview of the scale-invariant feature transform and local naïve Bayes nearest neighbor classifier.

3.1 Scale-invariant feature transform (SIFT)

SIFT was proposed by David Lowe in 1999 [22]. SIFT is an algorithm in computer vision to detect and describe the local features in images. SIFT [21,22], with its greatest characteristic—scale invariance, seems to be one of the best feature extractions in image processing. This model uses Difference of Gaussian (DoG) functioning as a kind of an improvement of Gauss–Laplace algorithm to achieve scale invariance. After making scale-space by DoG, SIFT compares each point with its adjacent 26 pixels, which is the sum of eight adjacent pixels in the same layer and 9 pixels in the upper and lower adjacent layers. The location and scale of this point are recorded whether it is minimum or maximum. Hence, SIFT not only detects all extreme points of DoG scale-space, but also locates them exactly. These extreme points are called keypoints. After that, low contrast and unstable edge points would be removed. For each keypoint, SIFT computes the gradient strength and direction of every neighborhood, then it votes in histogram for every neighborhood according to gradient directions. The summations are used as the gradient strengths of a keypoint. The maximal gradient strength is defined as the main direction of this keypoint. Then, a 16 $\times$ 16 region centered at the keypoint is chosen. After the region is chosen, it is divided into 4 $\times$ 4 sub-regions. The gradient strength in each sub-region is summed. An eight-dimensional vector is generated using eight directions in each sub-region. Thereby, SIFT gets a 128-dimensional feature description from 16 sub-regions.

The reason why we propose the use of SIFT as features of speech perception is that speech signals must be invariant under the transformation of the signals from speaker to speaker. Indeed, if they were not invariant in this way, some of the signals could only be produced by speakers with vocal tracts of a certain size. Actually, there are several invariances in speech sounds that are not quite so clearly dependent on the communication premise, and seem to have an influence on the nature of speech sound. Humans tend to speak at different speeds, with different loudness, or with varying dynamic range, when they are in different levels of stimulation. As a result, languages seem to adapt to this human feature by becoming more abstract in such a way that words are invariant under changes in tempo, pitch range, or emphasis. Presumably, this transformation of temporal stretching and compressing provides the invariance property of speech that plays an important role in the success of alphabets as descriptive devices for the sounds of human language. Alphabets are also invariant under changes in spacing and size of the letters. All these invariances of speech must be displayed in the spectrogram image of signals.

In addition, SIFT is an image descriptor for image-based matching and recognition [19,20]. These descriptors as well as related image descriptors are used for a large number of purposes in computer vision related to point matching between different views of a 3-D scene and view-based object recognition. The SIFT descriptor is invariant to translations, rotations and scaling transformations in the image domain and robust to moderate perspective transformations and illumination variations. Experimentally, the SIFT descriptor has been proven to be successful in practice for image matching and object recognition under real-world conditions [20]. For all these reasons, we would like to employ SIFT in our experiments.

3.2 Local naïve Bayes nearest neighbor

The LNBNN [20] was proposed by Sancho in 2012 to improve NBNN for image classification problems. The LNBNN can be described as follows. Supposing that we have to classify data into N classes C_1, C_2, …C_N, each of which has a number of training samples T_{i1}, T_{i2}, …, for i = 1,…N, and each sample is presented with a number of feature vectors. A feature vector is an m-dimension vector.

Firstly, LNBNN merges all feature vectors from the samples of all classes to build a kd-tree to speed up the nearest neighbor searching. In the classification phase, when a feature vector is queried, its $k + 1$ nearest neighbors are found. The $k + 1$ nearest neighbors are sorted in ascending order of distance to query the feature vector. Hence, the border distance is assigned by the distance of the $(k + 1)$th neighbor. LNBNN calculates minimum distance from the query feature vector to all feature vectors found in k nearest neighbors. The

minimum total distance determines the class of the query feature vector. In short, the LNBNN algorithm is described as follows:

Algorithm LocalNBNN (Q, k)
Input:
$Training\ set\ T = \{T_1, T_2, \ldots, T_N\}$
$T_i = \{d_{i_1}, d_{i_2}, \ldots, d_{i_{N_i}}\}$ and $d_{i_j} \in R^m\ \forall j = 1..N_i$
$C = \{C_1, C_2, .., C_N\}\ is\ a\ set\ of\ class\ label$
$Query\ Q = \{d_1, d_2, .., d_Q\}, d_i \in R^m\ \forall i = 1..Q$
Parameter k is number of nearest neighbor
Output: Class of Q
1: **for all** $d_i \in$ Q**do**
2: find $\{p_1, p_2, .., p_{k+1}\}$ *are* $k+1$ *nearest neighbors of* d_i
3: $dist_B = \|d_i - p_{k+1}\|^2$
4: **for all** classes C in the k nearest neighbors **do**
5: $dist_C = min_{\{p_j | Class(p_j) = C\}} \|d_i - p_j\|^2$
6: $totals[C] \leftarrow totals[C] + dist_C - dist_B$
7: **end for**
8: **end for**
9: **return** $\text{argmin}_C\ totals[C]$

4 Our approach

In this section, we describe the characteristics of speech signals' spectrogram images and how to extract SIFT [21,22] from these images. Then we describe our speech classification framework.

SIFT [21,22] is a feature extractor in image processing. Therefore, if we want to use SIFT features in speech classification, we must present speech signal in the form of its spectrogram. The spectrogram of speech is a visual representation of the spectrum of frequencies in a sound or other signal, as they vary with time or some other variable. The spectrogram was used to identify spoken words phonetically and to analyze the various calls of animals. After converting to a spectrogram, each speech signal is presented in a grayscale image. Then, we extract SIFT feature from the spectrogram images of all training and testing samples.

Figures 1, 2, 3, 4 and 5 are spectrogram images of some speech signals. Figure 1 shows the spectrograms images of the English alphabet A produced by four different speakers, while Fig. 2 describes the spectrogram images of A, B, C and D from the same speaker. Figure 3 illustrates the spectrogram images of five utterances spoken by five different speakers and Fig. 4 presents the spectrogram images of five utterances produced by a speaker. The last one, Fig. 5, is about some SIFT feature points extracted from A, B, C and D alphabet.

Figures 1 and 3 show that although speech signals are produced by different speakers, the speech signals of the same words tend to be similar in the spectrogram image. Besides, Figs. 2 and 4 show that signals of different words have different spectrogram images. It can be seen that the speech classification problems can turn into the image classification problems, which could inherit the results of computer vision field such as feature extraction. Moreover, both quantization and informative feature selection on a long-tail distribution will incur a large information loss. To reduce the loss discriminative feature, we need to maintain all features that are extracted from the training database, so each sample will have a different number of features. While most of the state-of-the-art classification approaches need input data to have the same size, the LNBNN classifier permits classifying data with different sizes of the feature vector. In addition, the LNBNN does not need to retrain old data when adding new data. This leads to the fact that LNBNN can learn incrementally.

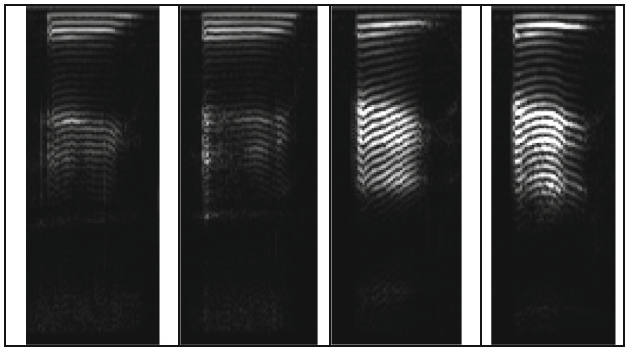

Fig. 1 Spectrogram images of the English alphabet *A* from four different speakers

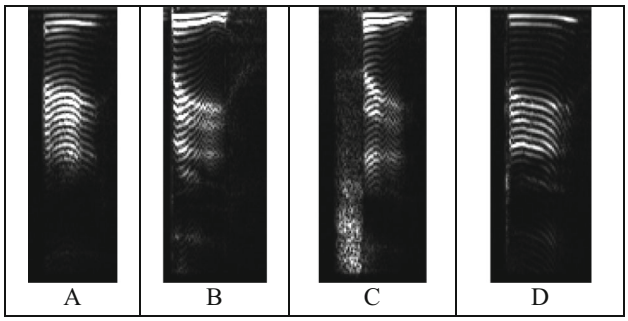

Fig. 2 Spectrogram images of the English alphabet A–D from the same speaker

In this paper, we propose an image base for the speech classification framework. Our framework has two phases: the preparing phase and the classification phase. In the preparing phase, firstly, all speech signals are converted to spectrogram image with windows size 10 ms and overlap 5 ms. Now, each speech signal is presented in a spectrogram image. Then, we use the SIFT extractor to extract the SIFT features from all training spectrogram images. Each spectrogram image is represented by a set of SIFT features. Next, we build a KD-TREE to boost up the KNN search in all SIFT features of training spectrogram images. In the classification phase, query speech is also converted to a spectrogram image and then the SIFT extracted from this. After that, we use SIFT feature of query spectrogram image and use LNBNN to clas-

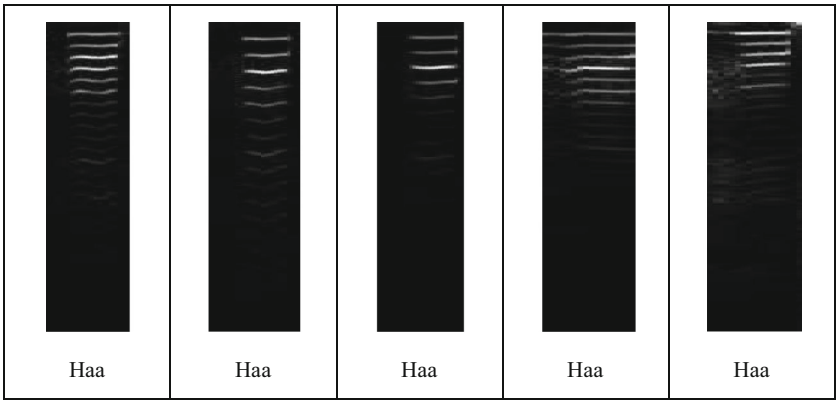

Fig. 3 Spectrogram images of five utterances in the JVPD database from five different speakers

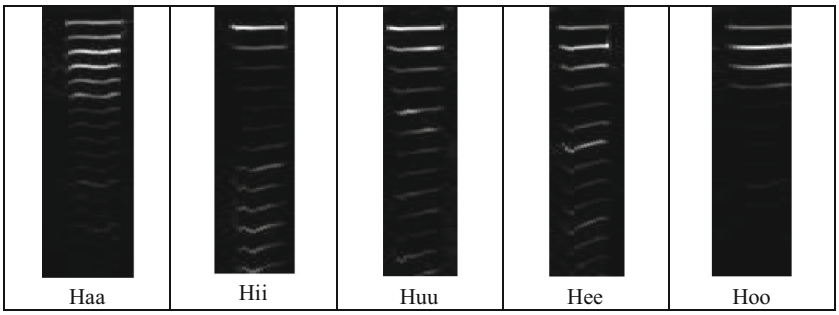

Fig. 4 Spectrogram images of five utterances in the JVPD from a speaker

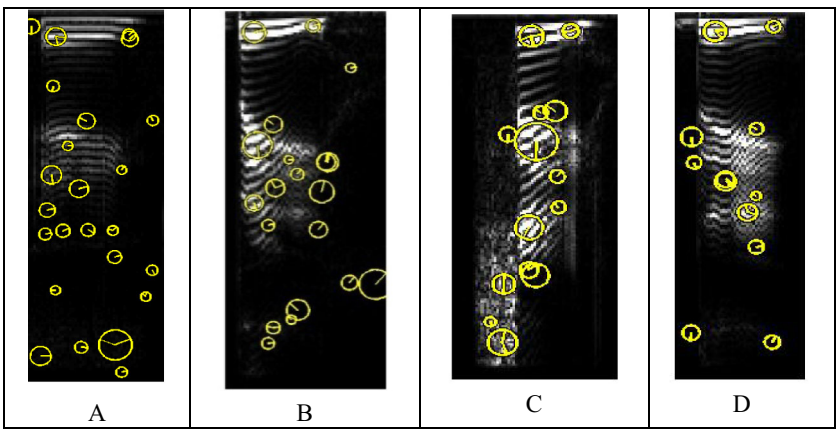

Fig. 5 SIFT features from some English alphabet *A*, *B*, *C* and *D*

sify this. Figure 6 describes our image-based framework for speech classification.

5 Experiments

In this section, we describe the six databases that are used in three experiments. In the first experiment, we compare the accuracy rate of LNBNN with SIFT and MFCC feature. In the second experiment, we compare the LNBNN classifier to some other popular machine learnings which are naïve Bayes, Bayesian network, support vector machine, random forest and decision tree analysis J48. In the third experiment, we evaluate the capacity of adding data in LNBNN after training. In the last experiment, we carry out two sub-experiments. First, we trained the model for all classes in the training database with small training samples and then add more samples for each class and evaluate the accuracy of the model when adding more data. Then, we perform using a small number of classes to train the model and then incremental update classes.

5.1 Experiment set

In our experiments, we use six databases, namely Isolated Letter (ISOLET) [3], English Isolated digits [51],

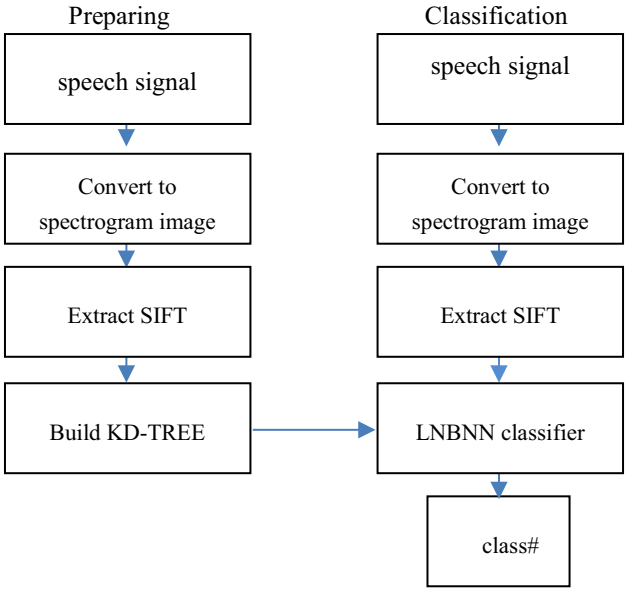

Fig. 6 Speech classification using LNBNN with SIFT

Vietnamese Places name [52], Vietnamese Digits, Tohoku University- Matsushita Isolated Word (TMW) [53], and Five Japanese Vowels of Males, Females, and Children Along with Relevant Physical Data (JVPD) [54].

The ISOLET database has 676 samples of spoken English letters. It was constructed by 26 speakers. The ISOLET database was divided into 20 training samples and 6 testing samples for each class. The Isolated digits database (0–9, o) has 454 samples for each class. This was divided into 266 training samples and 188 testing samples. The Places database has eight classes that were names of eight places (caphe, dung, karaoke, khachsan, khong, matxa, tramatm, trolai) in Vietnamese. Each class has 485 training samples and 50 test samples.

Vietnamese spoken digits database (Một, Hai, Ba, Bốn, Năm, Sáu, Bảy, Tám, Chín, Mười) was divided into 20 training samples and 5 testing sample for each class.

The TMW is Tohoku University-Matsushita Isolated Word Database. It has phonetically balanced 212 words that are spoken by 60 people (30 males and 30 females). This database was divided into the training and testing set. The training set has 40 samples and the test set has 20 samples.

JVPD is built on five Japanese Vowels of Males, Females, and Children which are along with Relevant Physical Data. The vowels are haa, hii, huu, hee, hoo. The speech data of men, women, and children ranging between 6 and 56 years of age were edited into files. Each utterance is spoken by 385 speakers (186 males and 199 females). The JVPD was divided into 269 training samples and 115 testing samples.

Table 1 Compare average correct classification rate of LNBNN with MFCC and SIFT features on six databases

Databases	SIFT	MFCC
ISOLET	0.73	0.34
English digits	0.96	0.94
Vietnamese places	0.95	0.39
Vietnamese digits	0.97	0.72
TMW	1.00	0.39
JVPD	0.97	0.53

5.2 Experiment with LNBNN in combination with SIFT and MFCC

In the first experiment, we used LNBNN in the classification step. In the feature extraction step, we used MFCC and SIFT to find suitable feature extraction for LNBNN in the speech classification approach. The experiment was deployed on all six above databases. Table 1 shows the average correct result for each database of classification.

Table 1 figures out the accuracy of LNBNN when using SIFT feature of spectrogram image higher than using MFCC. The highest different accuracy is in the TMW database. It is 39 % higher when using SIFT compared to using MFCC. The lowest different accuracy is in the English digits database. It is 2 % higher when using SIFT compared to using MFCC. This result shows that the SIFT feature is better for speech classification when using LNBNN.

Table 1 shows the average accuracy of LNBNN in combination with SIFT which is higher than combination with MFCC, while Figs. 7, 8, 9, 10 and 11 show that for each database. The accuracy of all classes using SIFT are mostly higher when using MFCC except in the English digit database. Two classes (three and six) have higher accuracy in the MFCC feature than using SIFT.

The first experiment shows that SIFT features are not only a good feature in the image processing field, but also better

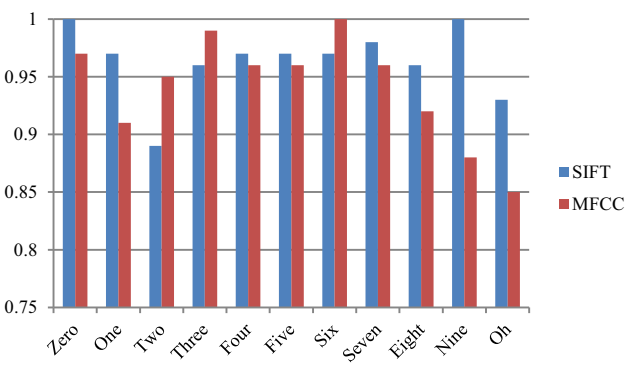

Fig. 7 Correct classification rate of LNBNN with MFCC and SIFT features on the English digits database

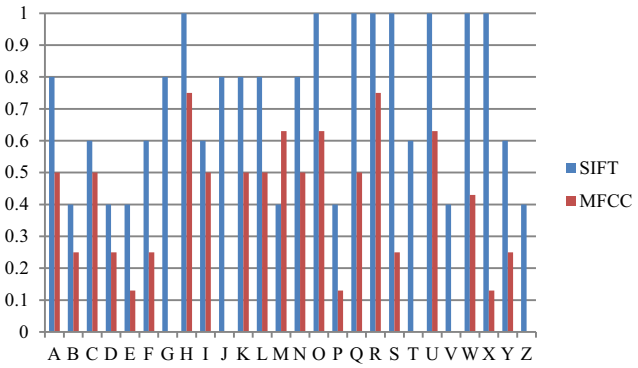

Fig. 8 Correct classification rate of LNBNN with MFCC and SIFT features on the ISOLET database

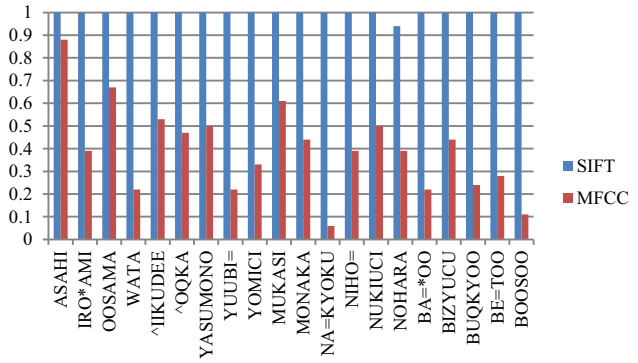

Fig. 9 Correct classification rate of LNBNN with MFCC and SIFT features on the TMW database

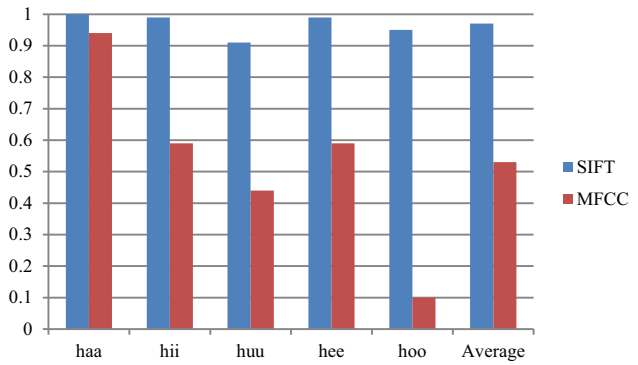

Fig. 10 Correct classification rate of LNBNN with MFCC and SIFT features on the JVPD database

than the familiar MFCC in speech perception when combined with the LNBNN classifier.

5.3 Experiments with LNBNN and other classifiers

In the second experiment, our objective was to evaluate the effectiveness of the LNBNN in speech classification by comparing this method with other approaches such as naïve Bayes, bayesian network, support vector machine (SVM), random forest and decision tree analysis J48 (Tree.J48).

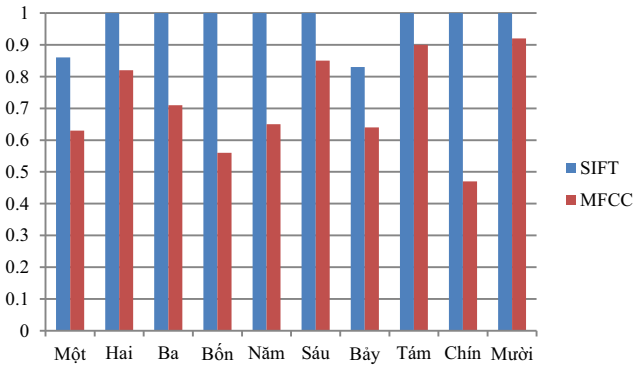

Fig. 11 Correct classification rate of LNBNN with MFCC and SIFT features on the Vietnamese digits database

In LNBNN, we use both SIFT and MFCC features extracted from spectrogram images of speech signals. Since other classifiers need to have the same dimension input data, we use the LBG algorithm to quantize features to the same dimension. SIFT features extracted from spectrogram images are quantized to 16 SIFT feature points; then these 16 feature points are converted to 128 $\times$ 16 dimension vectors for each sample. For MFCC features, we extract 18 MFCC coefficients from each speech signal, and then all MFCC coefficients are quantized to 16 feature vectors. After that, 16 MFCC vectors are converted to one 16 $\times$ 18 dimension vector.

Tables 2 and 3 show the average correct result for each classification approach on six database with MFCC and SIFT features.

Table 2 shows that LNBNN in combination with MFCCs does not have the highest correction rate. The LNBNN accuracy is slightly lower than other classifiers on all databases. In the ISOLET database, the LNBNN classifier has the lowest accuracy. In the Vietnamese places and TMW database, the accuracy of the LNBNN classifier is noticeably lower than most others methods.

In Table 3, the LNBNN with SIFT gives the highest correction rate compared to others for the speech classification problem. In the ISOLET database, LNBNN reaches 72.8 % correct classified samples, then random forest (64.4 %) and naïve Bayes (64.2 %). Especially, in English Isolated digits database, the LNBNN classifies correctly accounting for 96.2 %, then random forest reaches the rate of 70.7 %. The most apparent difference between LNBNN and other approaches is in JVPD and Vietnamese Places database. The LNBNN reaches the highest accuracy at 96.9 % and the second highest accuracy at 62.4 % in Random Forest on the former and LNBNN reaches the highest accuracy at 95.0 % and the second highest accuracy at 78.5 % in Random Forest on the later. Especially, in 20 first classes of TMW database, LNBNN in combination with SIFT classifies correctly all samples, while the second highest accuracy is 69.0 of Ran-

Table 2 Average correct classification rate with different methods with MFCC

Method	ISOLET	English digits	Vietnamese places	Vietnamese digits	TMW	JVPD
LNBNN	34.0	94.1	38.5	72.0	39.0	87.1
Naïve Bayes	64.2	98.6	67.6	42.4	44.6	44.5
Bayes Net	57.0	**99.5**	70.2	47.5	21.3	21.3
SVM	61.6	**99.5**	**78.0**	62.8	40.7	96.5
RandomForest	**64.4**	98.4	71.8	**73.5**	**56.7**	**97.2**
TreeJ48	38.1	90.2	53.8	42.4	15.2	82.7

Bold values indicate highest value in the column

Table 3 Average correct classification rate with different methods with SIFT

Method	ISOLET	English digits	Vietnamese places	Vietnamese digits	TMW	JVPD
LNBNN	**72.8**	**96.2**	**95.0**	**96.9**	**100.0**	**96.9**
Naïve Bayes	32.8	50.4	58.5	53.1	34.1	55.8
Bayes Net	20.6	57.2	70.5	47.7	33.1	60.8
SVM	3.8	11.3	12.5	14.6	8.5	35.2
Random forest	37.7	70.7	78.5	55.2	69.0	62.4
Tree J48	18.3	47.3	60.3	34.6	17.4	46.8

Bold values indicate highest value in the column

Table 4 Average correct classification rate on increment update samples training

Database	20 % Training samples	40 % Training samples	60 % Training samples	80 % Training samples	100 % Training samples
ISOLET	0.46	0.56	0.60	0.68	0.73
English digits	0.90	0.92	0.94	0.95	0.96
VN places	0.91	0.92	0.93	0.94	0.95
VN digits	0.27	0.72	0.71	0.82	0.97
TMW	0.92	0.93	0.98	0.99	1.00
JVPD	0.94	0.96	0.96	0.95	0.97

dom Forest and SVM has the lowest accuracy (8.5). Table 3 also shows that SVM has the lowest accuracy rate for all database when using quantized SIFT features.

5.4 Increment update training in LNBNN

One of the LNBNN classifier advantages is that LNBNN allows adding training samples without retraining the whole samples. In this section, we describe the capacity of increment update training data to the LNBNN. We carry out the experiment on adding more training samples for all classes after training. In this experiment, we divided training samples of each class into five portions of 20 %. We use increment update training data to model. First, we use 20 % of training data to build the model and test classification correction. Then, we add 20 % more training data, and up to 100 % training data are added to the model. Secondly, we perform an experiment on adding new classes after training the model. In this experiment, we divided training classes into five portions of 20 %. Thus, the experiment has five steps, starting with 20 % of classes for training and testing and ending with 100 % of training and testing classes added to the model. In both experiments, we use SIFT feature combination with the LNBNN classifier. Table 4 shows the average correction rate with increment update samples. Table 5 shows the average correction rate of increment update classes.

In Table 4, the VN digit has the highest difference correction rate in 20 % of training samples and 100 % of training samples. At the step when 20 % samples were added to the model, the VN digits reached 0.27 and at 100 % training samples step, 0.97. The second highest difference correction rate was on the ISOLET database and the lowest difference was on the JVPD database. Table 4 shows that, on most of the database, when training samples were added, the classification accuracy increased. However, in steps 2 and 3 on the JVPD database, the accuracy was not increased when training data were added, but it even decreased in step 4.

In Table 5, while most of the database has lower accuracy when more classes are added, the accuracy is almost not

Table 5 Average correct classification rate on increment update class training

Database	20 % Classes	40 % Classes	60 % Classes	80 % Classes	100 % Classes
ISOLET	0.55	0.64	0.60	0.60	0.73
English Digits	1.00	0.98	0.98	0.97	0.96
VN Places	1.00	0.97	0.95	0.94	0.95
VN Digits	1.00	0.97	0.98	0.96	0.97
TMW	1.00	1.00	1.00	1.00	1.00
JVPD	1.00	1.00	0.97	0.97	0.97

changed in the TMW and even increases in the ISOLET. This shows that when learning more knowledge, the model is more confusing in classification. However, this experiment has proved that the LNBNN allows adding more classes (new knowledge) without retraining the whole training samples.

6 Conclusion

In this paper, we have proposed an approach that uses LNBNN classifier in combination with SIFT features for the speech classification problem. The proposed approach allows adding training samples without retraining after the training phase. This saves training time which is suitable for big data. Another advantage of this method is that feature vectors do not need to be quantized. Such quality of input features is not diminished. It contributes to improve the quality of classification. As the above experiments figure out, the proposed approach performs well in a speech classification system. Classification of the speech signal based on combination of LNBNN and SIFT features gives better results than a combination of LNBNN and other features. In addition, SIFT features are more suitable for the LNBNN classifier than other classifiers.

In the future, we would like to improve SIFT-based LNBNN classifier to reduce the number of features. We also would like to modify SIFT to more suitable speech data.

References

1. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M.: Perception of speech code. Psychol. Rev. **74**, 431–461 (1967)
2. Liberman, A.M., Mattingly, I.G.: The motor theory of speech perception revised. Cognition **21**, 1–36 (1985)
3. Cole, R., Fanty, M.: ISOLET (Isolated Letter Speech Recognition), Department of Computer Science and Engineering, September 12 (1994)
4. Massaro, D.W.: Testing between the TRACE Model and the Fuzzy Logical Model of Speech perception. Cognitive Psychology, pp. 398–421 (1989)
5. McClelland, J.L., Elman, J.L.:The TRACE model of speech perception. Cognitive Psychology (1986)
6. Wilson, W., Marslen, M.: Functional parallelism in spoken word-recognition. Cognition **25**, 71–102 (1984)
7. Patel, I.: Speech recognition using HMM with MFCC—an analysis using frequency spectral decomposition technique. Signal & Image Proc Int J (SIPIJ). **1**(2) (2010)
8. Paul, D.B.: Speech Recognition Using Hidden Markov Models. Lincoln Lab. J. **3**(1) (1990)
9. Adam, T.B.: Spoken english alphabet recognition with mel frequency cepstral coefficients and back propagation neural networks. Int J Comput Appl. **42**(12), 0975–8887 (2012)
10. Salam, M.S.H., Mohamad, D., Salleh, S.: Malay isolated speech recognition using neural network: a work in finding number of hidden nodes and learning parameters. Int Arab J Info Technol **8**, 364–371 (2011)
11. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. In: IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 43–49 (1978)
12. Hinton, G., et al.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. In: IEEE Signal Process, pp. 82–97 (2012)
13. Abdel-Hamid, O., et al.: Convolutional neural networks for speech recognition in IEEE/ACM transactions on audio. Speech and language processing, October, USA (2014)
14. Hermansky: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. **87**(4), 1738–52 (1990)
15. Favero R.F.: Compound wavelets: wavelets for speech recognition. In: International symposium on time-frequency and time-scale analysis, pp. 600–603, (1994)
16. Jaitly, N., Hinton, G.: Learning a better representation of speech soundwaves using restricted boltzmann machines. In: Proc. of ICASSP, pp. 5884–5887 (2011)
17. Sainath T., Weiss, R., Senior, A., Wilson, W., Vinyals O.: Learning the Speech Front-end with Raw Waveform CLDNNs. In: Interspeech (2015)
18. Dimitri, P., Mathew, M.D., Ronan, C.: Analysis of CNN-based speech recognition system using raw speech as input. In: Interspeech (2015)
19. Boiman, O., Shechtman, E., Iran, M.: In defense of nearest-neighbor based image classification. In: CVPR (2008)
20. McCann, S., Lowe, D.G.: Local Naive Bayes nearest neighbor for image classification. In: CVPR (2012)
21. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: IJCV (2004)
22. Lowe, D.G.: Object recognition from local scale-invariant features. Proceedings of the international conference on computer vision **2**, 1150–1157 (1999)

23. Sakriani, S., Konstantin, M., Satoshi, N., Wolfgang, M.: Incorporating knowledge sources into statistical speech recognition.: Springer Science & Business Media (2009)
24. Sadaoki, F.: 50 years of Progress in speech and Speaker Recognition Research. vol. 1, no. 2, November (2005)
25. Davis K.H., Biddulph R., Balashek, S.: Automatic recognition of spoken digits. J. Acoust. Soc. Am, pp. 637–642 (1952)
26. Olson, H.F., Belar, H.: Phonetic typewriter. J. Acoust. Soc. Am. **28**(6), 1072–1081 (1996)
27. Fry D.B.: Theoretical aspects of mechanical speech recognition. J. Br. Inst. Radio Eng., pp. 211–299 (1959)
28. Rabiner, L.R., Levinson, S.E., Rosenberg, A.E., Wilpon, J.G.: Speaker independent recognition of isolated words using clustering techniques. IEEE Trans. Acoustics, Speech, Signal Proc (1979)
29. Sakoe, H.,: Two level DP matching—a dynamic programming based pattern matching algorithm for connected word recognition. IEEE Trans. Acoustics, Speech, Signal Proc., pp. 588–595 (1979)
30. Loizou, P.C., Spanias, A.S.: High-performance alphabet recognition. IEEE Trans. Speech Audio Proc. **4**, 430–445 (1996)
31. Cole, R., Fanty, M., Muthusamy, Y., Gopalakrishnan M.: Speaker-independent recognition of spoken english letters. In: International Joint Conference on Neural Networks (IJCNN), pp. 45–51 (1990)
32. Cole, R., Fanty, M.,: Spoken letter recognition. In: Presented at the Proceedings of the conference on advances in neural information processing systems Denver, Colorado, United States (1990)
33. Fanty, M., Cole, R.: Spoken Letter Recognition. In: Presented at the Proceedings of the conference on advances in neural information processing systems Denver, Colorado, United States (1990)
34. Karnjanadecha, M., Zahorian, S.A.: Signal modeling for high-performance robust isolated word recognition. IEEE Trans. Speech Audio Proc. **9**, 647–654 (2001)
35. Ibrahim, M.D., Ahmad, A.M., Smaon, D.F., Salam M.S.H.: Improved E-set recognition performance using time-expanded features. In: Presented at the second national conference on computer graphics and multimedia (CoGRAMM), Selangor, Malaysia (2004)
36. Jonathan, D., Da, T.H., Haizhou, L.: Spectrogram Image feature for sound event classification in mismatched conditions. In: IEEE Signal Processing letters, pp. 130–133 (2011)
37. Mohamed, A.R., Dahl, G.E., Hinton, G.E.: Deep belief networks for phone recognition. In: NIPS workshop on deep learning for speech recognition and related applications (2009)
38. Mohamed, A., Dahl, G., Hinton, G.: "Acoustic modeling using deep belief networks. In: IEEE Trans. Speech, & Language Proc, Audio (2012)
39. Mohamed, A., Hinton, G., Penn, G.: Understanding how deep belief networks perform acoustic modelling. In: Proc. ICASSP (2012)
40. Bocchieri, E., Dimitriadis, D.: Investigating deep neural network k based transforms of robust audio features for lvcsr. In: ICASSP (2013)
41. Tuske, Z., Golik, P., Schluter, R., Ney, H.: Acoustic modeling with deep neural networks using raw time signal for lvcsr. In: Interspeech (2014)
42. Palaz, D., Magimai, M., Collobert, R.: Convolutional neural networks-based continuous speech recognition using raw speech signal. In: ICASSP (2015)
43. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. J. Mach., pp. 103–130 (1997)
44. Behmo, R., Marcombes, P., Dalalyan, A., Prinet V.: Towards optimal naive bayes nearest neighbor. In: ECCV (2010)
45. Tuytelaars, T., Fritz, M., Saenko, K., Darrell, T.: The NBNN kernel. In: ICCV (2011)
46. Wang, J., Yang, J., Yu, K., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: CVPR (2010)
47. Liu, L., Wang, L., Liu, X.: In defense of soft-assignment coding. In: ICCV (2011)
48. Ma, C., O'Shaughnessy, D.: A perceptual study of source coding of Fourier phase and amplitude of the linear predictive coding residual of vowel sound. J. Acoust. Soc. Am. **95**(4), 2231–2239 (1994)
49. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of audio fingerprinting. J. VLSI Signal Proc. Syst. Signal Image Video Technol. **41**, 271–284 (2005)
50. Wang, A.L.C.: https://www.ee.columbia.edu/dpwe/papers/. Accessed 15 Nov 2015
51. https://catalog.ldc.upenn.edu/LDC2008S07. Accessed 15 Nov 2015
52. http://www.alovoice.vn/ai/du-lieu-tieng-noi-tieng-viet/. Accessed 15 Nov 2015
53. http://research.nii.ac.jp/src/en/TMW.html. Accessed 15 Nov 2015
54. http://research.nii.ac.jp/src/en/JVPD.html. Accessed 15 Nov 2015
55. Diehl, R.L., Lotto, A.J., Holt, L.L.: Speech perception. Annu. Rev. Psychol, pp. 149–179 (2004)
56. Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M.: Perception of the speech code. Psychol, Rev (1967)
57. Dah, G., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. In: IEEE Trans Speech, Lang Proc. Audio, USA (2012)