# Acoustic Feature Comparison for Different Speaking Rates

Abdolreza Sabzi Shahrebabaki(✉), Ali Shariq Imran, Negar Olfati, and Torbjørn Svendsen

Faculty of Information Technology and Electrical Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
{abdolreza.sabzi,ali.imran,negar.olfati,torbjorn.svendsen}@ntnu.no

**Abstract.** This paper investigates the effect of speaking rate variation on the task of frame classification. This task is indicative of the performance on phoneme and word recognition and is a first step towards designing voice-controlled interfaces. Different speaking rates cause different dynamics. For example, speaking rate variations will cause changes both in formant frequencies and in their transition tracks. A word spoken at normal speed gets recognized more often than the same word spoken by the same speaker at a much faster or slower pace, or vice-versa. It is thus imperative to design interfaces which take into account different speaking variabilities. To better incorporate speaker variability into digital devices, we study the effect of (a) feature selection and (b) the choice of network architecture on variable speaking rates. Four different features are evaluated on multiple configurations of Deep Neural Network (DNN) architectures. The findings show that log Filter-Bank Energies (FBE) outperformed the other acoustic features not only on normal speaking rate but for slow and fast speaking rates as well.

**Keywords:** Intrinsic variations · Speaking rate · Acoustic features
FBE · MFCC · DNN

## 1 Introduction

Speech is an integral component of how humans interact with digital devices these days - be it text to speech [1], keyword spotting [2], voice-controlled devices such as Apple's Siri[1], Microsoft's Cortana[2], and Google's Google Home[3], or smart gadgets. Speech, in contrast to gesture or touch-based systems, is a natural way of communicating with these devices. The accuracy of voice-controlled devices is highly dependent upon speech variability including speaking rate and speaking style. Speaking very fast or slow for instance, can easily lower the recognition accuracy in devices if they are not tuned for it.

[1] https://www.apple.com/ios/siri/.
[2] https://www.microsoft.com/en-us/windows/cortana.
[3] https://store.google.com/us/product/google_home?hl=en-US.


M. Kurosu (Ed.): HCI 2018, LNCS 10903, pp. 176–189, 2018.
https://doi.org/10.1007/978-3-319-91250-9_14

Late 90s saw a lot of research on the topic of the speaking rate effects on speech recognition performance e.g. [3–5]. The studies at that time verified that the speech recognition performance, e.g. Word Error Rate (WER) degrades significantly at a fast speaking rate [6,7]. Later, the performance degradation as a result of variations in speaking rate was also confirmed for speaker recognition in studies carried out in [8,9]. Attempts were made to figure out precisely how speaking rate effects Automatic Speech Recognition (ASR) performance by showing a direct correlation between local average Hidden Markov Model (HMM) score and local speech rate [10].

After almost twenty years, however, the research is still in a preliminary stage for most speech-based applications such as in the case of speaker authentication [11], where it is argued that an important reason for performance degradation is due to a distorted spectrum caused by variations in speaking rate [12], particularly for slow speaking rates. The rate at which people speak depends on many characteristics related to the speaker such as gender, age, and the psychological state they are in. For instance, a study presented in [13] showed that on average older people speak slowly compared to young ones and females talk slower than the males. Additionally, deviating speaking rate is often observed in our daily life. People usually speak fast when in hurry or angry or they may speak slow if they are tired, sad, or sick [14]. Speaking rate patterns also differs between the native speakers and non-native speakers. Research has shown that non-native speakers talk much slower compared to native speakers [15]. More recent studies also revealed that non-native speakers exhibit more variation in speaking rate [16]. On the other hand, this suprasegmental characteristic between native and non-native speakers in spontaneous speech suggests that non-native speakers are less variable than native speakers [17].

Speaking rate variability affects the mapping between the acoustic properties of speech and the linguistic interpretation of the utterances [18]. ASR systems employing supervised machine learning techniques and deep learning methods can efficiently learn the phonetic patterns. However, speaking rate variability can drastically decrease the performance of the ASR systems if they are not tuned for it. While listeners can naturally adapt to the changes in speaking rate and can maintain phonetic constancy, applying rate normalization in ASR systems for understanding phonetic patterns can be a challenging task.

This paper, therefore, investigates speaking rate variability from two perspectives: (a) which speech features perform best under variable speaking rate conditions? and (b) which DNN architecture obtains the highest accuracy on a frame classification task for speech recognition?

The remainder of the paper is structured as follows. Section 2 presents related work, followed by Sect. 3 where we describe DNN. Section 4 gives an overview of the experimental setup. Results and their analysis are presented in Sect. 5, while Sect. 6 concludes the paper with insight into the future work.

## 2 Related Work

Meyer et al. reported one of the earliest works that exploited the logatome speech database discussed in Sect. 4.1. They conducted a study on the performance of ASR to Human Speech Recognition (HSR) for several intrinsic variabilities such as speaking rate, speaking effort and dialect [19]. Their ASR model based on HMM uses a three-state-model for each phoneme. Describing each phoneme by a binary voicing and ternary features defining manner and place of articulation, they observed that misclassification of voicing and manner of articulation were the major causes for recognition errors. In a similar work [20], authors address reducing the ASR and human listeners gap while having a particular emphasis on intrinsic variations of speech. The work was further extended to use DNN as the ASR backend where phoneme confusion matrices obtained by ASR models for Mel Frequency Cepstral Coefficients (MFCC), FBE, and Perceptual Linear Prediction (PLP) features were compared against those obtained by human subjects [21]. FBE and PLP showed the highest correlation coefficient score between ASR and human subjects for various Signal to Noise Ratio (SNR) values.

Varghese and Mathew in [22] used a reservoir computing technique in their two-layered Recurrent Neural Network (RNN) for classifying 39 phoneme classes on the TIMIT database. They used the Relative spectral transformation Perceptual Linear Prediction (Rasta-PLP) and MFCC as features for frame level classification where MFCC performed marginally better than Rasta-PLP. A comparison of MFCC and supervised Isomap on the task of phoneme recognition is carried out in [23]. The authors also proposed a supervised manifold learning algorithm that outperforms the baseline MFCC and the supervised Isomap. Authors in [24] compared the performance of MFCC, PLP, and Rasta-PLP using fuzzy logic and Deep Belief Networks (DBN) on the African language phoneme classification. MFCC and Rasta-PLP results were far better than PLP while fuzzy logic classified consonants better than vowels with respect to DBN. A similar study related to phonetic analysis on Arabic speech is presented in [25] which compares six acoustic features that include Linear Predictive Coding (LPC), MFCC, PLP, FBE, Mel-filter bank coefficients (MELSPEC), and Linear Prediction Reflection Coefficients (LREFC). A five-state HMM is used to model each phoneme with a mixture of sixty-four Gaussian distributions. FBE achieved the highest accuracy while MELSPEC results were marginally behind followed by PLP and MFCC.

Comparison between different acoustic features have been addressed on different datasets and for various speech related activities, e.g. on digits [26], for event detection [27], and on emotional speech classification [28] among others. Not much can be found in the literature on how they perform under variable speaking rates. This paper, therefore, addresses the question of how these acoustic features compare to each other on different architecture combination, context size, and for different speaking rates.

## 3 DNN

A deep neural network is a term used for artificial neural network with several hidden layers. A Multi Layer Perceptron (MLP) consisting of at least two or more hidden layers is often used as a baseline DNN, unlike a vanilla network that consists of a single hidden layer. An MLP is a feedforward neural network in which all the neurons in one layer typically are fully connected to the neurons in the adjacent layer. The model uses two phases for estimating the weights, first in an unsupervised method the initial values for the weights are found and then in the second phase, the initialized weights are updated by a supervised technique called "backpropagation". The first phase is called pre-training, and the latter one is called fine-tuning. The training procedure of the DNN is described in the following subsections.

### 3.1 Pre-training

As we know initializing the weights when the network has multiple hidden layers is a bit challenging and will affect the convergence of the network weights. The main idea behind the pre-training is to find the initial weights which are estimated by fitting a generative DBN to the input data [29]. The DBN can be trained in a greedy layer by layer approach in which each pair of layers are considered as a Restricted Boltzmann Machine (RBM). An RBM has two layers, one of them contains visible nodes ($v = [v_1, v_2, ..., v_K]^T$) and the other one are hidden nodes ($h = [h_1, h_2, ..., h_L]^T$). There are different variations of RBM according to the data type available. When the input values are real-valued data, the Gaussian-Bernoulli RBMs are used, and when the input values are binary, the Bernoulli-Bernoulli RBMs are used. The difference between these two RBM is in the energy function definition. For the Bernoulli-Bernoulli RBMs, the energy function is defined as:

$$E(v,h) = -\sum_{k=1}^{K}\sum_{l=1}^{L} v_k h_l w_{kl} - \sum_{k=1}^{K} v_k a_k - \sum_{l=1}^{L} h_l b_l \tag{1}$$

where the $w_{kl}$ are the weights between the visible unit $v_k$ and the hidden unit $h_l$, $a_k$ is the bias for the visible unit $v_k$ and $b_l$ is the bias for the hidden unit $h_l$. The weights and biases are real-valued data, and the hidden and visible are binary-valued data. For the Gaussian-Bernoulli RBMs, the energy function is defined as:

$$E(v,h) = -\sum_{k=1}^{K}\sum_{l=1}^{L} \frac{v_k}{\sigma_k} h_l w_{kl} - \sum_{k=1}^{K} \frac{(v_k - a_k)^2}{2\sigma_k^2} - \sum_{l=1}^{L} h_l b_l \tag{2}$$

where the $\sigma_k$ is the standard deviation of the Gaussian noise for visible unit $v_k$ which is a real-valued unit. The joint probability of the the visible and hidden units is defined as follows:

$$p(v,h) = \frac{exp(-E(v,h))}{Z} \tag{3}$$

where $Z$ is the partition function which is sum over all values of $v, h$.

$$Z = \sum_{v,h} e^{-E(v,h)} \tag{4}$$

The weights, biases and the standard deviations are estimated during the training by maximizing the expected log probability, given in (5), of the visible units with the contrastive divergence (CD) algorithm [29].

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathbf{E}[\log\, p(v)] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ \mathbf{E}[\log \sum_{h} p(v,h)] \tag{5}$$

where $\boldsymbol{\theta}$ contains the weights, biases and standard deviations, $\hat{\boldsymbol{\theta}}$ is the estimated values for the parameters and $\mathbf{E}[.]$ is the expectation of the containing arguments. After training the first RBM on the input data which are visible units (v1), the hidden units (h1) are inferred. The inferred units are used as the visible units for the next RBM (v2 = h1) to estimate the hidden units (h2). For the number of hidden layers in DBN, RBMs are trained and stacked after each other. The Fig. 1 shows the stacked RBMs and the resulted DBN.

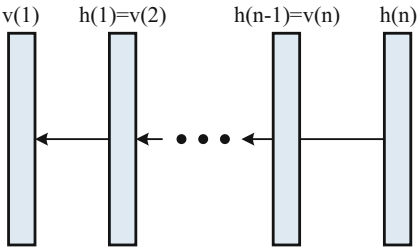


**Fig. 1.** Graphic model of DBN by stacking the RBMs.

### 3.2 Fine-Tuning

After unsupervised learning and estimating the initial values for the network parameters, supervised learning is performed by adding the labels as the output units on top of the DBN. The output weights are randomly initialized, and the cross-entropy cost function is considered to update the weights by minimizing the cross entropy between the estimated outputs and the labels by using the back-propagation algorithm. Because of the multiclass problem, a softmax function is considered at the output layer to estimate the probabilities of input samples classified to each class.

### 3.3 Architecture Configurations

Figure 2 shows the model architecture of a three-hidden layer DNN with 1024 neurons in each layer. For the experimentation, the following parameters of the DNN

are used: loss function: categorical cross entropy, learning rate: 0.01, optimizer: Stochastic Gradient Descent (SGD), activation function: sigmoid, batch size for training and prediction: 1024. A softmax function is used at the output layer.

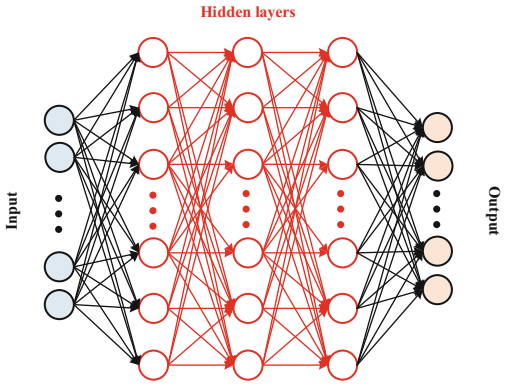


**Fig. 2.** Graphic model of DBN by stacking the RBMs.

## 4 Experimental Setup

### 4.1 Dataset

The Oldenburg Logatome (OLLO) corpus is used for the experiments. It is a speech database that contains simple non-sense combinations of consonants (C) and vowels (V), which are referred to as logatomes. 150 different CVCs and VCVs combination were spoken by 40 German and 10 French speakers. The VCVs are the combination of fourteen central consonants and five outer vowels. Also, eight consonants and ten vowels are combined to make the CVCs. In both combinations the outer phonemes are the same.

Four different dialects are covered by the German speakers: no dialect; Bavarian; East Frisian and East Phalian. The database contains logatome spoken at a normal pace, followed by variabilities such as, 'fast', 'slow', 'loud', 'soft' and 'questioning'. These variabilities can be grouped into three categories: (i) speaking rate (fast, slow and normal), (ii) speaking style (question and statement), and (iii) speaking effort (loud, soft and normal). Each of 150 logatomes has been repeated three times by each speaker. The same number of male and female speakers is used to record the database to cover the gender variabilities. The sampling frequency of the utterances is 16 kHz. OLLO has mostly been used for comparison between HSR and ASR [19,30]. We primarily chose to use this dataset for following reasons:

(a) Evaluating different variabilities and their effects on the ASR systems is possible by using this database.

(b) Also, OLLO may be useful in identifying how dialect and accent influence the speech recognition performance.

In the following experiments, the ten speakers with no dialect have been chosen. The variabilities fast, slow and normal are used.

### 4.2 Speech Features

This study uses the most popular acoustic features which include FBE, MFCC, LPC, and Line Spectral Frequencies (LSF). Features from multiple resolutions concerning both time and frequency domain are extracted, resulting in different frame shifts and different feature dimension respectively.

**FBE.** FBE features are extracted by a filter bank of 40 filters with uniform bandwidth on the mel frequency scale. The mel frequency scale closely resemble the frequency sensitivity of the human auditory system. FBE features are computed by taking a logarithm of the filterbank energies. MFCCs are then obtained by applying the DCT transformation on the FBE features. As a result of this transformation, the features become nearly uncorrelated. To preserve the information in both FBE and MFCC after using the DCT transform, all of the features are used and there is no dimensionality reduction.

**MFCC.** MFCCs are obtained by applying the DCT transformation on the FBE features. As a result of this transformation, the features become nearly uncorrelated. To preserve the information in both FBE and MFCC after using the DCT transform, all of the features are used and there is no dimensionality reduction.

**LPC.** According to the source-filter model of speech, the vocal tract acts as a filter on the excitation signal produced by lungs and vocal cords [31]. An all-pole filter is considered to model the vocal tract frequency response and the obtained coefficients as a result are the LPC features. These coefficients are extracted from a short time windowed signal to satisfy the quasi-stationarity of the modeled signal. The filter order is chosen as 40 to match the dimensionality of the FBE and MFCC input features, higher than for typical LP analysis of speech.

**LSF.** Line spectral frequencies or Line Spectral Pairs (LSP) is another variant of LPC features which is less sensitive to quantization noise compared to the LPC features. LSF order is kept as same as the LPC order which is 40.

### 4.3 Context Dependent Feature Representation

The input to the DNN is a context-dependent feature vector $\boldsymbol{x}_c(n)$ which is computed by considering the frames on the left and right side of the current frame $x(n)$. $M$ is the number of preceding and following frames that are concatenated with the current frame $x(n)$ to constitute the DNN input vector. The left and

right context size can vary, but for these experiments, both are kept same. The concatenated input vector shown in (6) is of size $D \times (2M + 1)$, where $D$ is the feature vector dimension and $M$ is the context size. In the experiments, four different context sizes ($M = 3, 5, 7$ and 10) are considered to assess the effect of context on the frame classification accuracy.

$$\boldsymbol{x}_c(n) = [\boldsymbol{x}(n-M)^T, \ldots, \boldsymbol{x}(n)^T, \ldots, \boldsymbol{x}(n+M)^T]^T, \tag{6}$$

where the $T$ is transpose operator.

## 5 Results and Discussion

To evaluate the effect of different speaking rate on the frame classification performance, several experiments are conducted. Four different feature types are extracted from slow, normal and fast speaking rate by using 25 ms frame length and 10 ms frame shift. The frame classification task is performed by sequentially selecting one speaker for testing and the remaining speakers for training the classifier which implies a speaker independent frame classification task. The number of phone classes is 24. DNN is chosen as the classifier for this task. The training procedure of DNN is according to Sect. 3. Several experiments were conducted by varying the number of neurons (128, 256, 512, 1024) and size of the hidden layers (2, 3, 4, 5) for all feature sets.

Our findings revealed that 512 and 1024 neurons in each layer have the highest accuracy rate for 3 and 4 hidden layers network. This paper, therefore, presents the results for 3 and 4 layers architecture only having 512, and 1024 neurons. The Tables 1, 2, 3 and 4 show the frame accuracy rate for the training and test data with normal speaking style. The performance of FBE is higher than the other feature types whereas LPC has the worst performance. By looking at the effect of context size, we see that by moving from context size $M = 3$ to the higher values, the performance increases significantly. From context size $M = 5$ to $M = 10$ there is not that much increase in the performance for LSF and MFCC, somehow the performance is saturated, but for the FBE it has almost one percent better accuracy rate.

**Table 1.** Frame accuracy rate for different features and different structures for normal speaking style and context size $M = 3$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | Train | Test | Train | Test | Train | Test | Train | Test |
| 3layers-512nodes | 86.75 | 60.14 | 86.50 | 66.56 | 91.22 | 68.39 | 89.45 | 76.31 |
| 3layers-1024nodes | 86.50 | 60.03 | 86.16 | 66.35 | 91.23 | 69.86 | 89.47 | 76.48 |
| 4layers-512nodes | 85.27 | 59.06 | 85.93 | 66.52 | 91.46 | 68.95 | 89.36 | 75.91 |
| 4layers-1024nodes | 84.56 | 59.11 | 84.73 | 65.96 | 86.16 | 69.59 | 88.55 | 76.02 |

**Table 2.** Frame accuracy rate for different features and different structures for normal speaking style and context size $M = 5$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | Train | Test | Train | Test | Train | Test | Train | Test |
| 3layers-512nodes | 87.60 | 62.04 | 89.14 | 72.82 | 94.53 | 72.99 | 93.53 | 82.70 |
| 3layers-1024nodes | 90.62 | 64.30 | 91.52 | 73.73 | 94.99 | 74.29 | 93.74 | 82.95 |
| 4layers-512nodes | 87.55 | 61.55 | 89.88 | 72.56 | 93.52 | 72.39 | 93.34 | 82.56 |
| 4layers-1024nodes | 87.87 | 63.24 | 90.43 | 73.14 | 93.86 | 74.35 | 93.54 | 82.73 |

**Table 3.** Frame accuracy rate for different features and different structures for normal speaking style and context size $M = 7$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | Train | Test | Train | Test | Train | Test | Train | Test |
| 3layers-512nodes | 91.24 | 62.09 | 93.00 | 72.16 | 97.55 | 74.17 | 94.12 | 82.31 |
| 3layers-1024nodes | 92.13 | 63.19 | 92.80 | 72.72 | 96.83 | 73.66 | 94.24 | 82.13 |
| 4layers-512nodes | 89.98 | 60.81 | 93.03 | 71.92 | 96.10 | 72.82 | 93.99 | 82.36 |
| 4layers-1024nodes | 89.56 | 61.34 | 92.86 | 72.02 | 95.92 | 73.12 | 94.17 | 82.08 |

**Table 4.** Frame accuracy rate for different features and different structures for normal speaking style and context size $M = 10$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | Train | Test | Train | Test | Train | Test | Train | Test |
| 3layers-512nodes | 92.99 | 62.51 | 95.08 | 73.53 | 97.28 | 73.81 | 95.68 | 83.58 |
| 3layers-1024nodes | 93.57 | 64.13 | 94.31 | 73.75 | 97.38 | 74.88 | 95.59 | 83.44 |
| 4layers-512nodes | 86.54 | 60.29 | 93.44 | 72.76 | 95.43 | 72.68 | 94.64 | 83.05 |
| 4layers-1024nodes | 90.79 | 62.37 | 93.58 | 73.46 | 97.01 | 74.75 | 94.81 | 83.22 |

In order to examine whether speaking rate would have different impact on consonant and vowel recognition, we have looked at the average accuracy of the frame classification for vowels and consonants, respectively. In addition, the misclassifications were broken down into two separate categories for both vowels and consonants: confusions within the broad class (e.g. a consonant misclassified as another consonant) and confusions between the classes (e.g. a vowel misclassified as a consonant). Figure 3 shows these performance measures for test data with

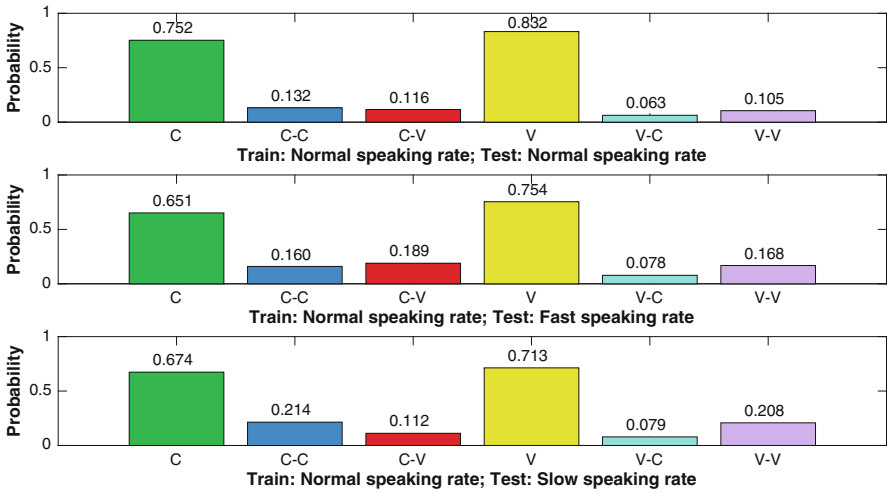


**Fig. 3.** Rates for correct and incorrect classification of consonants and vowels. The green bar shows the correct classification rate for consonants, blue bar shows the misclassification of a consonant as another consonant, red bar shows the confusion of the consonants with vowels. The yellow bar shows the correct classification rate of vowels, cyan bar shows the misclassification of a vowel as a consonant and the purple bar shows the confusion of the vowels with another vowel. (Color figure online)

| | /a/ | /ɛ/ | /ɪ/ | /ɔ/ | /ʊ/ | /aː/ | /e/ | /i/ | /o/ | /u/ |
|---|---|---|---|---|---|---|---|---|---|---|
| /a/ | **0.58** | 0 | 0 | 0.04 | 0 | 0.29 | 0 | 0 | 0 | 0 |
| /ɛ/ | 0 | **0.79** | 0.06 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ɪ/ | 0 | 0.06 | **0.54** | 0 | 0 | 0 | 0.11 | 0.16 | 0 | 0 |
| /ɔ/ | 0 | 0 | 0 | **0.73** | 0.06 | 0.04 | 0 | 0 | 0.08 | 0 |
| /ʊ/ | 0 | 0 | 0 | 0.06 | **0.52** | 0 | 0 | 0 | 0.13 | 0.19 |
| /aː/ | 0 | 0 | 0 | 0 | 0 | **0.88** | 0 | 0 | 0 | 0 |
| /e/ | 0 | 0 | 0.06 | 0 | 0 | 0 | **0.62** | 0.22 | 0 | 0 |
| /i/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0.91** | 0 | 0 |
| /o/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0.76** | 0.16 |
| /u/ | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0.11 | **0.81** |

(a) Slow speaking rate

| | /a/ | /ɛ/ | /ɪ/ | /ɔ/ | /ʊ/ | /aː/ | /e/ | /i/ | /o/ | /u/ |
|---|---|---|---|---|---|---|---|---|---|---|
| /a/ | **0.83** | 0 | 0 | 0 | 0 | 0.05 | 0 | 0 | 0 | 0 |
| /ɛ/ | 0 | **0.86** | 0.04 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| /ɪ/ | 0 | 0.06 | **0.79** | 0 | 0 | 0 | 0.03 | 0.05 | 0 | 0 |
| /ɔ/ | 0.03 | 0 | 0 | **0.81** | 0.05 | 0 | 0 | 0 | 0 | 0 |
| /ʊ/ | 0 | 0 | 0 | 0.05 | **0.78** | 0 | 0 | 0 | 0 | 0.02 |
| /aː/ | 0.03 | 0 | 0 | 0 | 0 | **0.88** | 0 | 0 | 0 | 0 |
| /e/ | 0 | 0 | 0.1 | 0 | 0 | 0 | **0.72** | 0.1 | 0 | 0 |
| /i/ | 0 | 0 | 0.03 | 0 | 0 | 0 | 0 | **0.91** | 0 | 0 |
| /o/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **0.85** | 0.09 |
| /u/ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.08 | **0.84** |

(b) Normal speaking rate

**Fig. 4.** Vowel part of the confusion matrix for different speaking rate

different speaking rates using FBE features as the input vector with context size $M = 10$ to a DNN with 1024 hidden nodes in each of the three hidden layers. By considering the normal speaking rate as the reference point, we can see that the true classification rate of the consonants in fast speaking rate is the lowest one, and it is confused more with vowels. Also the true classification rate of the vowels in slow speaking rate is the lowest one and the vowels confusion increased.

For justification of these claims, we look at the confusion matrices for slow and normal speaking rates. In Fig. 4 it can be easily found that the slow speaking rate causes the confusion between long and short vowels more than the normal speaking rate. It is worth mentioning here that the summation of the rows of confusion matrices in Fig. 4 are not equal to one, because the confusion with the consonants are not shown here.

The results in the Tables 5, 6, 7 and 8 are from the same networks as in the previous experiments. The only difference here is that the networks are trained on the 'normal' speaking rates while the performance is evaluated on the slow and fast speaking test data sets. In these experiments, the FBE results are superior among the other feature types. The context size does not have any effect on the LPC results, but for the other features the performance increases moderately. FBE has the higher performance even with the smaller context size. For the FBE the accuracy rate for fast speaking style is always better than slow speaking rate, but for the other feature types, the slow speaking rate has better performance than the fast speaking rate within the longer context size.

**Table 5.** Frame accuracy rate for fast (V1) and slow (V2) speaking styles on the networks trained on normal speaking style, context size $M = 3$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | V1 | V2 | V1 | V2 | V1 | V2 | V1 | V2 |
| 3layers-512nodes | 57.83 | 58.02 | 63.22 | 63.34 | 64.77 | 64.57 | 72.71 | 70.47 |
| 3layers-1024nodes | 57.61 | 57.91 | 63.12 | 63.31 | 66.34 | 65.06 | 72.92 | 71.07 |
| 4layers-512nodes | 56.87 | 56.82 | 62.13 | 63.48 | 65.06 | 64.58 | 72.32 | 69.68 |
| 4layers-1024nodes | 56.60 | 57.19 | 62.56 | 62.99 | 65.50 | 64.84 | 72.51 | 70.58 |

**Table 6.** Frame accuracy rate for fast (V1) and slow (V2) speaking styles on the networks trained on normal speaking style, context size $M = 5$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | V1 | V2 | V1 | V2 | V1 | V2 | V1 | V2 |
| 3layers-512nodes | 57.48 | 58.53 | 65.06 | 65.88 | 67.08 | 67.77 | 75.48 | 73.63 |
| 3layers-1024nodes | 58.11 | 58.93 | 65.76 | 66.19 | 67.37 | 67.48 | 75.65 | 73.39 |
| 4layers-512nodes | 56.91 | 58.32 | 64.73 | 65.27 | 66.44 | 67.06 | 74.89 | 73.12 |
| 4layers-1024nodes | 56.87 | 57.69 | 65.25 | 65.64 | 66.14 | 65.65 | 74.88 | 72.73 |

**Table 7.** Frame accuracy rate for fast (V1) and slow (V2) speaking styles on the networks trained on normal speaking style, context size $M = 7$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | V1 | V2 | V1 | V2 | V1 | V2 | V1 | V2 |
| 3layers-512nodes | 57.86 | 58.76 | 66.01 | 67.21 | 67.87 | 68.85 | 75.81 | 74.42 |
| 3layers-1024nodes | 56.71 | 57.92 | 66.33 | 67.34 | 68.09 | 68.13 | 75.82 | 74.67 |
| 4layers-512nodes | 58.24 | 59.61 | 65.56 | 67.66 | 66.86 | 67.42 | 75.93 | 74.69 |
| 4layers-1024nodes | 56.92 | 58.19 | 66.09 | 66.81 | 67.82 | 67.81 | 75.81 | 74.37 |

**Table 8.** Frame accuracy rate for fast (V1) and slow (V2) speaking styles on the networks trained on normal speaking style, context size $M = 10$

| | Features | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| DNN | LPC | | LSF | | MFCC | | FBE | |
| Structure | V1 | V2 | V1 | V2 | V1 | V2 | V1 | V2 |
| 3layers-512nodes | 56.98 | 59.20 | 67.73 | 68.37 | 67.17 | 68.03 | 75.52 | 74.92 |
| 3layers-1024nodes | 58.55 | 60.23 | 66.61 | 68.85 | 68.19 | 69.16 | 75.79 | 74.91 |
| 4layers-512nodes | 54.92 | 57.16 | 65.23 | 67.88 | 66.05 | 67.75 | 75.24 | 74.29 |
| 4layers-1024nodes | 57.08 | 58.99 | 66.47 | 68.43 | 67.90 | 68.97 | 75.59 | 74.85 |

## 6 Conclusion

The paper provides a comparative analysis of acoustic features for LPC, LSF, MFCC, and FBE trained using DNN on slow, fast, and normal speaking utterances. Different combinations of DNN architectures by varying the number of layers and the number of nodes in each layer are tested. Three layers architecture with 512 and 1024 nodes in each layer performed well. Further experiments by varying the context window size for each feature are performed. Our initial findings revealed that on different context sizes, FBE achieved the highest frame classification accuracy for the normal speaking style. A similar trend was observed when the classifier was trained on the normal speaking rate and tested on slow and fast speaking rate. It was also observed that the bigger the context window, the better the classification accuracy.

Future work should focus on evaluating different deep learning classifiers such as those suited for predicting time series data, e.g., long short-term memory to see the effect of phoneme recognition on variable speaking rates.

## References


1. Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 20, pp. 1713–1724 (2013)
2. Chen, G., Parada, C., Heigold, G.: Small-footprint keyword spotting using deep neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4087–4091 (2014)
3. Martinez, F., Tapias, D., Alvarez, J., Leon, P.: Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition. In: Fifth European Conference on Speech Communication and Technology (EUROSPEECH), pp. 469–472 (1997)
4. Brondsted, T., Madsen, J.P.: Analysis of speaking rate variations in stress-timed languages. In: Fifth European Conference on Speech Communication and Technology (EUROSPEECH), pp. 481–484 (1997)
5. Martinez, J.F., Tapias, D., Alvarez, I.: Toward speech rate independence in large vocabulary continuous speech recognition. In: International Conference on Signal and Speech Processing, pp. 725–728 (1998)
6. Pfau, T., Ruske, G.: Creating hidden markov models for fast speech. In: Fifth International Conference on Spoken Language Processing, pp. 205–208 (1998)
7. Wrede, B., Fink, G.A., Sagerer, G.: An investigation of modelling aspects for rate-dependent speech recognition. In: Proceedings of the INTERSPEECH, pp. 2527–2530 (2001)
8. Xu, M., Zhang, L., Wang, L.: Database collection for study on speech variation robust speaker recognition. In: Proceedings of the O-COCOSDA (2008)
9. Grimaldi, M., Cummins, F.: Speech style and speaker recognition: a case study. In: Tenth Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 920–923 (2009)
10. Faltlhauser, R., Ruske, G., Thomae, M.: Towards the question: why has speaking rate such an impact on speech recognition performance? In: Seventh International Conference on Spoken Language Processing (ICSLP), pp. 2429–2432 (2002)
11. Rozi, A., Li, L., Wang, D., Zheng, T.F.: Feature transformation for speaker verification under speaking rate mismatch condition. In: Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4. IEEE (2016)
12. Zeng, X., Yin, S., Wang, D.: Learning speech rate in speech recognition. In: Proceedings of the INTERSPEECH, pp. 528–532 (2015)
13. Yuan, J., Liberman, M., Cieri, C.: Towards an integrated understanding of speaking rate in conversation. In: Proceedings of the INTERSPEECH, pp. 541–544 (2006)
14. Rao, K.S., Koolagudi, S.G.: Robust emotion recognition using speaking rate features. In: Robust Emotion Recognition using Spectral and Prosodic Features. Springer Briefs in Electrical and Computer Engineering, pp. 85–94. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6360-3_5
15. Guion, S.G., Flege, J.E., Liu, S.H., Yeni-Komshian, G.H.: Age of learning effects on the duration of sentences produced in a second language. Appl. Psycholinguistics **21**, 205–228 (2000)
16. Baese-Berk, M.M., Morrill, T.H.: Speaking rate consistency in native and non-native speakers of English. J. Acoust. Soc. Am. **138**(3), EL223–EL228 (2015)
17. Morrill, T., Baese-Berk, M., Bradlow, A.: Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of English. In: Proceedings of the International Conference on Speech Prosody, pp. 1119–1123 (2016)

18. Francis, A.L., Nusbaum, H.C.: Paying attention to speaking rate. In: Fourth International Conference on Spoken Language (ICSLP), pp. 1537–1540. IEEE, October 1996
19. Meyer, B.T., Wesker, T., Brand, T., Mertins, A., Kollmeier, B.: A human-machine comparison in speech recognition based on a logatome corpus. In: Workshop on Speech Recognition and Intrinsic Variation, pp. 95–101 (2006)
20. Meyer, B.T., Brand, T., Kollmeier, B.: Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. J. Acous. Soc. Am. **129**(1), 388–403 (2011)
21. Exter, M., Meyer, B.T.: DNN-based automatic speech recognition as a model for human phoneme perception. In: Proceedings of the INTERSPEECH, pp. 615–619 (2016)
22. Varghese, D., Mathew, D.: Phoneme classification using reservoirs with MFCC and Rasta-PLP features. In: Computer Communication and Informatics (ICCCI), pp. 1–6. IEEE (2016)
23. Yang, J., Cao, T., Sun, X., Huang, S., Huan, L.: Phoneme classification based on supervised manifold learning. In: Robotics and Applications (ISRA), pp. 931–934. IEEE (2012)
24. Laleye, F.A., Ezin, E.C., Motamed, C.: Adaptive decision-level fusion for Fongbe phoneme classification using fuzzy logic and deep belief networks. In: Informatics in Control, Automation and Robotics (ICINCO), pp. 15–24. IEEE (2015)
25. Meftah, A., Alotaibi, Y.A., Selouani, S.A.: A comparative study of different speech features for arabic phonemes classification. In: Modelling Symposium (EMS), pp. 47–52. IEEE (2016)
26. Bharali, S.S., Kalita, S.K.: A comparative study of different features for isolated spoken word recognition using HMM with reference to Assamese language. Int. J. Speech Technol. **18**(4), 673–684 (2015)
27. Kiktova, E., Lojka, M., Pleva, M., Juhar, J., Cizmar, A.: Comparison of different feature types for acoustic event detection system. In: Dziech, A., Czyżewski, A. (eds.) MCSS 2013. CCIS, vol. 368, pp. 288–297. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38559-9_25
28. Sukhummek, P., Kasuriya, S., Theeramunkong, T., Wutiwiwatchai, C., Kunieda, H.: Feature selection experiments on emotional speech classification. In: Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1–4. IEEE (2015)
29. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comput. **18**(7), 1527–1554 (2006)
30. Meyer, B., Brand, T., Kollmeier, B.: Phoneme confusions in human and automatic speech recognition. In: Proceedings of the INTERSPEECH, pp. 1485–1488 (2007)
31. Markel, J.D., Gray, A.J.: Linear Prediction of Speech, vol. 12. Springer, Heidelberg (2013)