MACHINE LEARNING APPROACH FOR VOICED/UNVOICED/SILENCE SPEECH SEGMENT DETECTION

Conference Paper · March 2013

CITATIONS
0

READS
288

2 authors, including:

Mohammed Abebe Yimer
Arba Minch University
2 PUBLICATIONS **0** CITATIONS

SEE PROFILE

MACHINE LEARNING APPROACH FOR VOICED/UNVOICED/SILENCE SPEECH SEGMENT DETECTION

Mohammed Abebe, moshethio@gmail.com[1]
Sebsibie Hailemariam (PhD), sebsibe2004@yahoo.com[2]
[1]Department of Computer Science & IT, Arba Minch University, Arba Minch, Ethiopia
[2]Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia

ABSTRACT

In this study a supervised method of voiced/unvoiced/silence speech segment detection is developed on Amharic speech corpus. A total of 900 Amharic sentences are collected which are acoustically tagged with three possible acoustic tags (voiced, unvoiced and silence). A speech corpus is prepared by recording these 900 sentences. Then the speech signal is segmented at 25 ms frame length with no overlapping. The researchers extracted frame energy, number of zero crossings (ZCR), Linear Prediction Coefficients (LPC), Mel Frequency Cepstral Coefficient (MFCC), Mel Frequency Cepstral Coefficient-Principal Component Analysis (MFCC-PCA) features from each frame and combined in six different forms.

The features were further processed to generate feature inventory for classification purpose by assigning the dependent variable (class label) which is the three values (voiced, unvoiced or silence). Using this feature, we train five different classification models: two rule-based (Decision Tables and JRip), two decision trees (J48 and simple CART) and a neural network or MLP (with one and two hidden layers). The MLP with one hidden layer shows the highest performance using the combined feature (MFCC, Energy and ZCR) with a performance of **89.33%**. Further, the researchers attempted to tune the parameters on MLP by changing the frame size, the learning rate, number of hidden layers and the number of neurons per hidden layer of the MLP. Finally, the experiments show that best performance with the selected classifier model and feature vector is achieved on 35 ms frame size with an accuracy of **89.69**%.

Keywords: Voicing Detection, Voiced, Unvoiced, Silence, Machine Learning, Speech Segment Detection

1. INTRODUCTION

The knowledge of acoustic speech feature in particular voiced or unvoiced segment plays an important role in many speech analysis-synthesis systems. Thus the issue of voicing detection (Voiced/Unvoiced/Silence) algorithms (VDAs) has been one of the topics most analyzed in the field of speech processing research during the last three decades (Beritelli, F., *et al*, 2009).

Voiced/Unvoiced/Silence (VUS) speech segment detection is a method of assigning and labeling a specific speech category (voiced/unvoiced/silence) to a speech segment. An accurate classification of a speech segment with voicing detection is often used as a prerequisite for developing other higher level and efficient applications of speech processing systems such as speech coding, speech analysis, speech synthesis, automatic speech recognition, noise suppression and enhancement, speaker identification, and the recognition of speech pathologies.

Voiced/Unvoiced detection involves identifying the regions of speech when there is significant glottal activity (i.e., the vibration of vocal folds). Such regions of speech are generally referred to as voiced speech (Dhananjaya, N., & B.Yegnanarayana, 2010). Voiced speeches include all vowels and some consonants, such as /m/, /n/, /l/, /w/, /b/, /d/, /g/, /v/, and /z/. Unvoiced speeches are produced by a turbulent air flow crossing some constriction in the vocal tract, without vibration of the vocal cords. Unvoiced sounds include consonants like /p/, /t/, /k/, /f/. Silence is produced as a result of air pressure emanating from lung without any constriction along the path in the speech production system (Ladefoged, P., 2001).

2. APPROACHES FOR VUS DETECTION

So far many voicing detection researches have been done and different approaches have been used for (VUS) classification, where the well-known ones are rule-based, statistical and neural network approach. The rule based approach as its name indicates relies on rules which are either handcrafted or machine learned rules. The rules are the important elements

for annotating speech segments in the rule-based approach (Bachu, R. G., et al, 2010).

The statistical based detector relies on the statistical property of speech signals. Such statistical property can be distributional probability of speech signals with tags which can be obtained during the training phase of the system. The most common statistical based voiced/unvoiced/silence speech segment detection method is HMM. This approach trains its parameters from observational feature data and does detection at later stage with or without adaptation (Tatarinov, J., & Pollak, P,). Finally, neural network (NN) detectors use acoustic features of speech signals that can be extracted from the speech waveform or spectrogram to classify speech categories.

In this study, a procedure is developed for making the VUS classification using various machine learning approaches constrained on various parameters such as speech segment frame length, features used and others as appropriate to the specific classifier.

3. GENERAL DESIGN OF VUS MODEL

The overall model of the training and classification processes for the VUS classifier is shown in Figure 1.

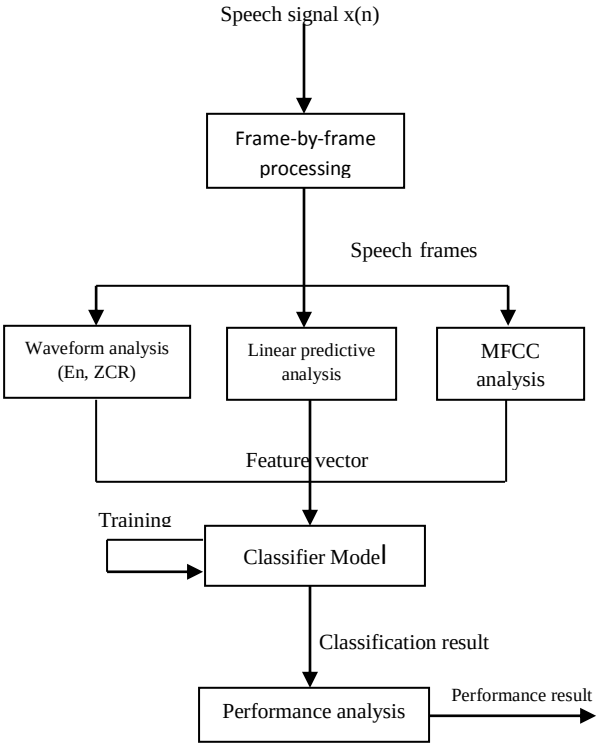

Figure 1: Design of VUS model

As shown in Figure 1, at first stage, the speech signal is segmented into frames without overlapping. In the frame-by-frame processing stage, the speech signal is segmented into a non-overlapping frame of samples. The researchers consider 4 different frame sizes for the research purpose: 20, 25, 30 and 35 milliseconds. From each frame we extracted frame energy, number of zero crossing (ZCR), Linear Prediction Coefficient (LPC feature vector of 13 dimension), Mel Frequency Cepstral Coefficient (MFCC feature vector of 13 dimension), Mel Frequency Cepstral Coefficient-Principal Component Analysis (MFCC-PCA feature vector of 13 dimension) features.

4. CORPUS PREPARATION

In this research, an optimal text selection technique is used to prepare text corpus from various sources. These documents are used as data sources to get phonetically rich and balanced collections of sentences. Amharic Bibles, health news, political news, sport news, economy news, penal code, Federal Negarit Gazeta and Amharic fiction named "Fiker Eskemekaber" are data sources used for text corpus preparation. This text corpus contains 900 Amharic sentences.

These sentences are recorded by a male Amharic language speaker person aged 29. Then the speech data is split into two sets: training (600 Sentences) and test set (300 sentences) using systematic random sampling. Both the training and testing data set is manually segmented for the voicing (voiced/unvoiced/silence).

5. FEATURE EXTRACTION

The dataset is further processed to prepare features to train and test the different classifiers. From each frames, six different combinations of features were generated. These are: (1) Energy with ZCR; (2) LPC; (3) Energy, ZCR with LPC; (4) MFCC; (5) Energy, ZCR with MFCC and (6) Energy, ZCR with MFCC-PCA. The complete feature to train the classifier is designed by taking the feature extracted from each frame as values of the independent variable and the segment label that is assigned manually (voiced/unvoiced/silence) as the dependent variable value (class label). As a result, 41, 118 feature vectors were prepared for testing purpose and 370, 062 feature vectors for training purpose.

6. SELECTION OF CLASSIFIERS

An extensive search for an optimal classifier is undertaken to empirically select a classifier that has a simple architecture and reasonably high classification

performance. Five different types of classifiers are tested namely: two rule-based (Decision Tables and JRip), two decision trees (J48 and simple CART) and a neural network (single and double hidden layers with different number of neurons on the hidden layer). Initial experiments were conducted to select the best classifier.

The number of dependent variable depends on the features considered as stated in the feature selection procedure. Table 1 shows the details of the classifiers performance on the different feature combination. Only the best two classifiers are indicated on the table.

Table 1: Best classifiers for the different combination of feature vectors

No	*Feature Used*	*1st Best Performance*	*2nd Best Performance*	*1st Best Model*	*2nd Best Model*
1	Energy & ZCR	74.33%	74.32%	MLP10, J48	DT, J48
2	LPC	87.67%	84.55%	J48	CART
3	Energy, ZCR & LPC	85.11%	84.77%	MLP10-5	MLP10
4	MFCC	89.08%	88.88%	MLP10	MLP10-5
5	Energy, ZCR & MFCC	89.33%	89.01%	MLP10	MLP10-5
6	Energy, ZCR & MFCC-PCA	88.58%	88.57%	MLP10-5	MLP10

(Note: MLP10 refers to MLP with one hidden layer composed of 10 neurons; MLP10-5 refers to MLP with two hidden layers composed of 10 and 5 neurons on the first and second hidden layer respectively)

From all the models tested, the neural network model shows a good classification performance on the different sets of feature combinations except for the second type feature category where J48 perform much better followed by CART. Moreover, the feature vector combinations having energy, zero-crossing rate and MFCC performs best on the neural network architecture. Furthermore, the 5th feature combination (Energy, ZCR & MFCC) is found to be best with an initial performance 89.33% with MLP10 (single hidden layer with 10 neurons at the hidden layer) classifier on the 35 millisecond frame size. This MLP10 classifier is composed of 15 input layer neurons that matched the dimension of the feature vector (13 MFCC coefficients, Energy and zero-crossing rate). There were 3 neurons on the output layer that matched the dimension of the classes (voiced, unvoiced and silence).

7. PARAMETER TUNING AND EXPERIMENTAL ANALYSIS

An extensive search for an optimal classifier is undertaken to empirically select a classifier that has a simple architecture and reasonably high classification performance. In this section, the single hidden layer MLP network which is proved to provide better performance is analyzed for its performance improvement through parameter tuning. These parameters are the number of hidden layers, number of neurons in the hidden layers, the effect of the frame size, and learning rate parameter α. Experiment on the optimal number of neurons in the hidden layer is made by changing from 2 neurons to 30 neurons. α value is taken at α=0.3 and 0.5.

These experiments are conducted by tuning the parameters of the MLP classifier using only feature vector five above due to the fact that it shows relatively high classification performance over the other feature vectors tested. The performance obtained for these experiments is summarized on Table 2.

Table 2: Analysis of MLP at different frame size and α value

Frame Size	*Model Type*	*Best Performance* with α =0.3	*Best Performance* with α =0.5
20 millisecond	MLP30	89.27%	89.15%
25 millisecond	MLP30	89.53%	89.36%
30 millisecond	MLP30	89.21%	89.09%
35 millisecond	MLP25	89.69%	89.25%

As it can be seen from the performance summary on Table 2, the performance of all the classifiers decreases as the learning rate is increased from 0.3 to 0.5. Finally, a neural network model with single hidden layer having 25 neurons on the hidden layer at the learning rate α=0.3 is selected as it gives a good classification performance on 35 ms speech segment.

Detail analysis of the performance of the classifier reveal out of the 41,118 test samples, 36,879 (89.6906%) were correctly classified and 4,239 (10.3094%) were incorrectly classified. The true positive rate (TP rate), False positive rate (FP rate), Precision, recall, F-Measure and the area under the

ROC curve information of the three distinct classes are indicated in Table 3.

Table 3: Performance detail by Class (MLP25)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Silence	0.778	0.011	0.827	0.778	0.802	0.978
Voiced	0.956	0.225	0.924	0.956	0.94	0.936
Unvoiced	0.709	0.043	0.799	0.709	0.751	0.919
Weighted Avg.	0.897	0.176	0.894	0.897	0.895	0.935

According to Table 3, the TP rate of voiced class is higher than the other two classes which shows that the proportion of correctly classified into their class is higher for voiced segments than others. At the same time the FP rate of voiced class is higher than other which shows more samples are classified into this class wrongly than other classes.

Table 4 shows the confusion matrix of the three classes. As shown in the Table, most of the voiced samples are wrongly classified into unvoiced than silence. Similarly, most unvoiced samples are classified into voiced. However, miss in silence ratio is almost uniform in both voiced and unvoiced classes. From this, one could observe the confusability of the voiced and unvoiced sample than silences. Moreover, the performance of the classifier varies for the different sound classes with a higher performance for voiced sounds followed by silence sounds and unvoiced sounds for the given testing set.

Table 4: Confusion Matrix (MLP25)

		Predicted			
		Silence	**Voiced**	**Unvoiced**	**Class Performance**
Actual Class	Silence	2055	304	281	77.84%
	Voiced	194	29173	1142	95.62%
	Unvoiced	237	2081	5651	70.91%

8. CONCLUSION

In this research, different classifiers with different types of frame based feature vector on a frame with different frame length were considered to build voiced/unvoiced/silence speech segment detection. Huge amount of corpus were developed and split randomly for testing and training purpose. The search space for the best model on these constraints was systematically handled.

Firstly, we tried to identify the best classifier and we found that multilayer perceptron with single hidden layer using the combined features (ZCR, Energy and MFCC) performed better than the remaining four classification models. The researcher then indentified the appropriate frame size, learning rate parameter for MLP weight adjustment. Accordingly, a multilayer perceptron with a single hidden layer having 25 neurons on the hidden layer on a feature vector generated at frame size of 35 milliseconds were identified as the optimal classification model. Finally, the classification performance of the identified MLP was analyzed.

9. RECOMMENDATION

There are lots of researches in natural language processing that can be done for different languages in Ethiopia. The same thing holds true for Amharic language. Therefore, to assist researchers, it will be of great paramount if a standard speech corpus for Amharic language is developed that will be available for NLP researchers in Amharic language.

Finally this research work suggests the following items as a future work:

- Comparative study of three different approaches (HMM based, rule based, ANN based classifiers with more training and testing data)
- Extending this work by training in large corpus and using balanced tag-sets for the different sound categories
- Comparison of hybrid approaches
- Experimenting phoneme identification following voiced/unvoiced/silence recognition
- Modeling and researching ANN based speech recognition system.

10. REFERENCES

Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2010). Voiced/Unvoiced Decision for Speech

Signals Based on Zero-Crossing Rate and Energy. USA.

Beritelli, F., Casale, S., Russo, A., & Serrano, S. (2009). Adaptive V/UV Speech Detection Based on Characterization of Background Noise.

Dhananjaya, N., & B.Yegnanarayana. (2010). Voiced/Nonvoiced detection based on robustness of voiced epochs IEEE Signal Processing Letters.

Ladefoged, P. (2001). A course in Phonetics (4 ed.). USA: Heinle & Heinle, a Division of Thomson Learning Inc.

Tatarinov, J., & Pollak, P. Hidden Markov Models in voice activity detection. Czech Republic

Qi, Y., & Hunt, B. R. (1993). Voiced-Unvoiced-Silence Classification of Speech Using Hybrid Features and Network Classifier. IEEE Transactions on Speech and Audio Processing, Vol. 1, NO. 2.