journal homepage: www.sabauni.net/ojs

S.J.I.T.N
Saba Journal of Information Technology and Networking

# Article

## EFFECT OF ACCENT ON IMPROVING SPEECH RECOGNITION SYSTEMS

**Amer A. Sallam[1], Nashwan A. Al-Khulaidi, Mogeeb A. Saeed**

**Taiz University**

**Article info**

**Article history:**

Accepted: Sep. 2018

**Keywords:**
Speech Recognition,
Speaker Identification,
Human Computer,
Interaction,
Signal Processing,
K-means.

**Abstract**

Speech is the output of a time varying excitation excited by a time varying system. It generates pulses with fundamental frequencies F0. This time varying impulse is trained as one of the features, and characterized by fundamental frequency F0 and its formant frequencies. These features vary from one speaker to another and from a gender to another one as well. In this paper the accent issues in continuous speech recognition system are considered. Variations in F0 and formant frequencies are the main features that characterize variation in a speaker. The variation becomes considerably less within a speaker, medium within the same accent and very high among a different accent. This variation in information can be exploited to recognize gender type and to improve performance of speech recognition systems through customizing separate models based on gender type information.

Five sentences are selected for training. Each of the sentences are spoken and recorded by 5 female speakers and 5 male speakers. The speech corpus will be preprocessed to identify the voiced and unvoiced region. The voiced region is the only region which carries information about F0. From each voiced segment, F0 is computed. Each forms the feature space labeled with the speaker identification: i.e., male or female. This information is used to parameterize the model for male and female. The K-means algorithm is used during training as well as testing. Testing is conducted in two ways: speaker dependent testing and speaker independent testing. SPHINX-III software by Carnegie Mellon University has been used to measure the accuracy of speech recognition of data taking in to account the case of gender separation which has been used in this research.

* Corresponding author: Amer A. Sallam
E-mail: amer.sallam@taiz.edu.ye

## 1-Introduction:

As speech recognition is a complex task it is still difficult to find the complete solution, because every human being has his/her own different characteristics of voice and accent, this in itself has become one of the main problems in the field of speech recognition [1], [2]. It is worth pointing out the fact that this research investigates an approach for identifying genders from spoken data and builds separate models for accent that can enhance the performance of speech recognition by reducing the search space in lexicon[3], [4],[5], [6].

Studies in gender classification using voice shall give insights to how humans process voice information[7], [8]. What may be important is that gender information is conveyed by the F0, type of sound, and the size of vocal tract. It can be assumed that modeling the vocal tract using Linear Prediction Coding (LPC) and Spectrum information[9],[10]. Furthermore, small errors in gender classification can be allowed; as sometimes it is even hard for a person to identify the gender of a speaker. This study has showed that a feature-based system with a trained decision tree can successfully classify male and female voices automatically. However, their studies have been most concentrated on age separation. In the context of speech recognition, accent and gender separation can improve the performance of speech recognition by limiting the search space to speakers from the same accent and gender.

In the gender classification accuracy[11], [12], it has been noticed that perhaps the most important variability across speakers besides gender is a role played by the accents. Therefore, it is probable that any recognition system attempting to be robust to a wide variety of speakers and languages should make some effort to account for different accents that the system might encounter.

In this paper, one approach is to use gender dependent features, such as the pitch and formants. The pitch information was used in [13], [14]for the problem of gender separation. However, fundamental frequency and formant frequencies estimation both rely considerably on the speech quality and accent. Although the quality of speech used in this study was not free from noise, we tried to improve the gender model by using K-means algorithm to get high accuracy[15]–[18].

## 2. Implementation

Gender separation is the process of separating the speaker's gender type directly from the acoustic information. This is possible using gender invariant feature separation and learning from variations within gender and accent. Furthermore, Speech is produced as a result of convolution of excitations of the vocal cord with the vocal tract system all coupled with the nasal tract.

It is decided to make a practical study of accent issues in continuous speech recognition systems, trying to find out the most correct results by implementing autocorrelation techniques. This practical study was carried out though several stages that will be described as follows.

**Stage 1: Data Collection**

**Training Data:**

5 female and 5 male speaker subjects are selected to record 5 properly selected sets of sentences. The selected recorded sentences for the training purpose are: welcome, where is Mike? I believe you are fine! Have fun with him. Thanks to God.

**Testing Data:**

For the testing purpose, 5 females and 5 males have been selected to speak one sentence and each has been recorded. In fact, every speaker has been given a sentence different from other

speakers. It has to be accounted that the group of testing is actually independent from that of the training data.

**Stage 2: Feature Extraction**

This stage illustrates the processes of features extraction that can identify the gender type on the basis of individual information included in speech waves through extracting the features, viz, fundamental frequency F0 from the collected training data; making separate models depending on the type of features and optimal parameters for each gender. The fundamental frequency shows high variations from one speaker to another. The variation becomes higher when the comparison is among speakers of different gender.

This process is represented in Figure 1. below. Technically, when the voice characteristics of utterance are checked, there will be a wide range of probabilities that exist for parametrically representing the speech signal for the gender separation task.

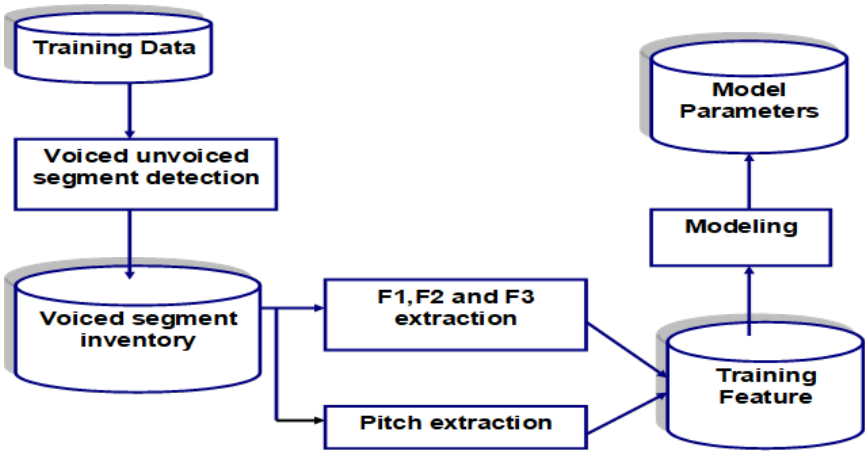


Fig1. Extraction of Features

**Stage 3: Feature Matching and Decision Making**

In this stage, we extracted the same feature type for testing data as done during the training stage. Then we applied the concept of pattern recognition to classify objects of interest into one of the desired gender types. The objects of interest are called sequences of vectors that are extracted from an input speech. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching as shown in Figure 2

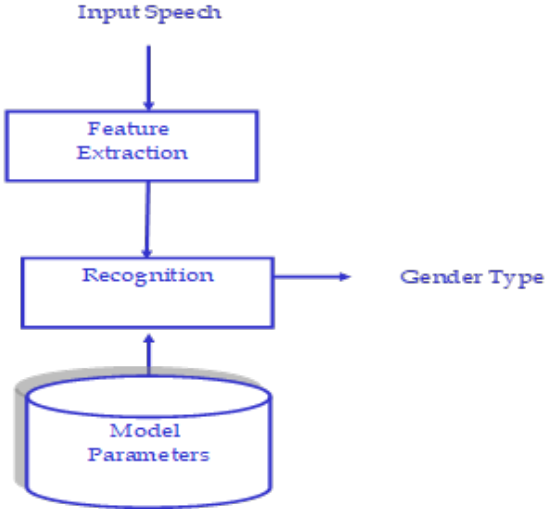


Fig2. Matching of Features

The K-means approach is used because of its flexibility and ability to give high recognition accuracy. K-means can be simply defined as a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its center; called a code word. The collection of all code words is called a codebook. The gender separation System will compare the codebooks of the tested speaker with the codebooks of the trained speaker. In other words, during recognition, among the models, the best model that maximizes the joint probability of the given observation will be selected as a recognized model. The best matching result will be the desired gender type and this can be verified as the decision making logic.

Finally, the gender type which is modeled by the recognized model will be given as an output of our system.

## 3. Data Analysis and Observation

In this section, the data analysis of the work is discussed. Figure 3. shows feature extraction (F0) in which the utterance welcome of both genders is analyzed showing that female on the right and male on the left. It shows that female speakers have higher pitch than male. Regarding the technique of feature extraction, each subfigure has its description as below in Figure 3.

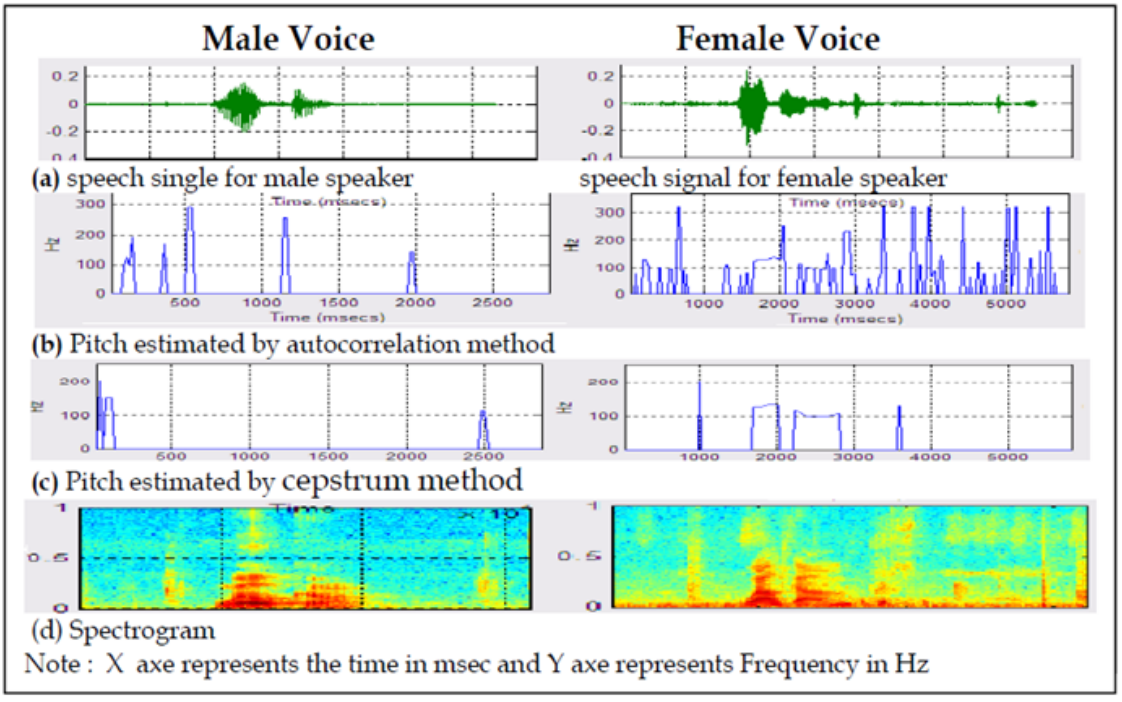


Fig3. Feature Extraction & Experiment Results

**Speech Recognition Accuracy Results:**

Speech recognition accuracy has been measured under Sphinx System and the results are as follows (see Table 1):

When the data is mixed from both genders, the accuracy resulted in 58%, and we consider this as a low result of accuracy. But when we separate the gender type, an increase of accuracy is obtained. This appears obviously through the results achieved by the same gender (females) which is 84%, while the accuracy results of the same gender (males) is 78%. In the step followed, we test the accuracy within the same gender and same speaker and the accuracy resulted is 100%.

In addition, for the training data of males and testing data of females, the accuracy of speech recognition resulted in 45%, while training data of females and testing data of males resulted in 34%.

Moreover, the results of same accents (Indian Accent) appear to be somewhat different for both genders. Their accuracy is 70 %. And when they are separated the accuracy of males is 72 % whereas the females is 90 %.

To sum up, we conclude that the gender sep-

aration and accent plays an important role in increasing the rate of accuracy of the results of speech recognition.

| Accuracy | Testing Data | | Training Data | |
|---|---|---|---|---|
| % | Female | Male | Female | Male |
| | | | | |
| Different accent | | | | |
| 58 % | □ | □ | □ | □ |
| 84 % | □ | | □ | |
| 78 % | | □ | | □ |
| 34 % | □ | | | □ |
| 45 % | | □ | □ | |
| Same accent | | | | |
| 70 % | □ | □ | □ | □ |
| 90 % | □ | | □ | |
| 72 % | | □ | | □ |
| Same speaker | | | | |
| 100 % | | □ | | □ |
| 100 % | □ | | □ | □ |
| | | | | |

Table 1: Speech Recognition Accuracy of Different Accent, Same Accent and Same Speaker

## 4.Conclusions

Speech is considered as the essential form of communication between humans. It plays a central role in the interaction between human and machine, and between machine and machine. The automatic speech recognition is aimed to extract the sequence of spoken words from a recorded speech signal and so it does not include the task of speech understanding, which can be seen as an even more elaborate problem. Because of the fact that the goal of speech recognition is still far away from the optimal solution for higher accuracy, the accent and the gender separation system has been proposed to enhance the performance of speech recognition through building separate gender accent models; by limiting search from whole space of acoustic models that can further lead to improve the accuracy of speech recognition. Although the speech data used in this study are collected from different nationalities with different accents (Arabs, Russians, Americans and Indians), it is recorded in different sampling rate and channels. High accuracy of gender recognition is obtained by using the technique which has been mentioned so far. However, when applying K-means algorithm as pattern recognition for the extraction of features, the results of the experiments for estimated pitch value through Autocorrelation and spectrum, have shown 100 % accuracy. Consequently, we conclude that pitch features are more suitable and strongly advised to distinguish the gender type. Moreover, we have noticed that the accent variability among speakers plays a crucial role in speech recognition accuracy besides gender feature. Therefore, it seems that any recognition system attempting to be robust to a wide variety of speakers and languages should make some effort to account for different accents that the system might encounter.

In future work, we are planning to expand the range of evaluation set. We are also planning to investigate the performance of the system to make it applicable, usable as well as useful for the study purpose in speaker-separation and speaker-verification.

## References:

[1] C. Qin and L. Zhang, “Deep neural network based feature extraction using convex-nonnegative matrix factorization for low-resource speech recognition,” 2016 IEEE Information Technology, Networking, Electronic and Automation Control Con-

ference. pp. 1082–1086, 2016.
[2] H. Aronowitz, "Speaker recognition using matched filters," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5555–5559, 2016.
[3] S. G. Koolagudi, B. K. Vishwanath, M. Akshatha, and Y. V. S. Murthy, "Performance Analysis of LPC and MFCC Features in Voice Conversion Using Artificial Neural Networks BT - Proceedings of the International Conference on Data Engineering and Communication Technology: ICDECT 2016, Volume 2," C. S. Satapathy, V. Bhateja, and A. Joshi, Eds. Singapore: Springer Singapore, 2017, pp. 275–280.
[4] R. Soorajkumar, G. N. Girish, P. B. Ramteke, S. S. Joshi, and S. G. Koolagudi, "Text-Independent Automatic Accent Identification System for Kannada Language," in Proceedings of the International Conference on Data Engineering and Communication Technology, 2017, pp. 411–418.
[5] H. Leach, K. Watson, and K. Gnevsheva, "Perceptual dialectology in northern England: Accent recognition, geographical proximity and cultural prominence," J. Socioling., vol. 20, no. 2, pp. 192–211, 2016.
[6] P. Adank, M. L. Noordzij, and P. Hagoort, "The role of planum temporale in processing accent variation in spoken language comprehension," Hum. Brain Mapp., vol. 33, no. 2, pp. 360–372, 2012.
[7] S. Garg and M. C. Trivedi, "Gender Classification by Facial Feature Extraction Using Topographic Independent Component Analysis," in Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, 2016, pp. 397–409.
[8] S. F. Abdullah, A. Rahman, Z. A. Abas, and W. H. M. Saad, "Development of a Fingerprint Gender Classification Algorithm Using Fingerprint Global Features," Development, vol. 7, no. 6, 2016.
[9] S. G. Koolagudi, B. K. Vishwanath, M. Akshatha, and Y. V. S. Murthy, "Performance Analysis of LPC and MFCC Features in Voice Conversion Using Artificial Neural Networks," in Proceedings of the International Conference on Data Engineering and Communication Technology: ICDECT 2016, Volume 2, C. S. Satapathy, V. Bhateja, and A. Joshi, Eds. Singapore: Springer Singapore, 2017, pp. 275–280.
[10] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), 2016, pp. 493–497.
[11] I. Heazlewood, J. Walsh, M. Climstein, J. Kettunen, K. Adams, and M. DeBeliso, "A comparison of classification accuracy for gender using neural networks multilayer perceptron (MLP), radial basis function (RBF) procedures compared to discriminant function analysis and logistic regression based on nine sports psychological constructs to measure motivations to participate in masters sports competing at the 2009 world masters games," in Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS), 2016, pp. 93–101.
[12] M. Castrillón-Santana, J. Lorenzo-Navarro, and E. Ramón-Balmaseda, "Multi-scale score level fusion of local descriptors for gender classification in the wild," Multimed. Tools Appl., pp. 1–17, 2016.
[13] T. Białaszewski and Z. Kowalczuk, "Solving Highly-Dimensional Multi-Objective Optimization Problems by Means of Genetic Gender," in Advanced and Intelligent Computations in Diagnosis and Control, Springer, 2016, pp. 317–329.
[14] P. S. Rathore and B. K. Joshi, "Gender recognition using FB series expansion and SVM," Signal Processing, Computing and Control (ISPCC), 2015 International Conference on. pp. 411–414, 2015.
[15] R. J. Kuo and P. S. Li, "Taiwanese export trade forecasting using firefly algorithm based K-means algorithm and SVR with wavelet transform," Comput. Ind. Eng., vol. 99, pp. 153–161, 2016.
[16] J. Lei, T. Jiang, K. Wu, H. Du, G. Zhu, and Z. Wang, "Robust K-means algorithm with automatically splitting and merging clusters and its applications for surveillance data," Multimed. Tools Appl., pp. 1–17, 2016.
[17] Q. Liu, W. Fu, J. Qin, W. X. Zheng, and H.

Gao, “Distributed k-means algorithm for sensor networks based on multi-agent consensus theory,” in 2016 IEEE International Conference on Industrial Technology (ICIT), 2016, pp. 2114–2119.

[18] G. Wu, Y. Xu, D. Wu, M. Ragupathy, Y. Mo, and C. Chu, “Flip-flop clustering by weighted K-means algorithm,” in Proceedings of the 53rd Annual Design Automation Conference, 2016, p. 82.