Arabic speech recognition using phase parameter

Tariq A. Hassan  
Al-Mustansiriyah University, Collage of Education, Computer Science Department, Iraq  
E-mail: tariqmono2016@uommustansiriyah.edu.iq

Duraid T. Salim  
Al-Mustansiriyah University, Collage of Basic education, Computer Science Department, Iraq  
E-mail: draid.edbs@uomstansiriyah.edu.iq

Abstract. In this paper, the main idea is to investigate the ability of the information presented in the speech signal phase in contributing in the speech recognition system. The database that we used is Arabic words (20 words) pronounced by 10 speakers in different times. The system adopts single frequency signals strategy by using filter bank in feature extraction stage. Gaussian model is used as a classifier. The results shows that will be absolutely error to ignore the phase parameter, especially in some system like MFCC and PLP models, which used only spectrum information. No comparisons are made; the paper is just examining the phase parameter under different circumstance, which include different Filterbank channels and different database styles (clean and noisy utterance).

Keywords: Speech recognition, biometrics, instantaneous components, feature extraction, Gaussian model.

1. Introduction

Speech signal, which is a collection of many components presented in one signal, can carry much information about the speech and speaker alike [1]. In terms of speech recognition, the focus will on those parameters in speech that hold unchanged regardless of the sentence that the speaker going to say [2]. On the other hand, speech recognition technique is the way to recognise the words (or sentences) said by different speakers in different situation. In this case, the parameters that we are going to deal with will be completely different from those in the case of speaker recognition. For example, in speaker recognition, the goal of work is to select features that minimize within-speaker variability and, at the same time, maximize between-speaker variability [3]. In term of speech recognition, the goal is to find features that minimize within-words variability and, at the same time, maximize between-words variability. In cases, speech and speaker recognition, humans are extremely brilliant in identifying both speech and speaker from the spoken words. This ability to do so encourage us to ask what are the parameters that used by human brain which make it recognise the speaker and the spoken words instantaneously. In this, many models have been proposed. Some are based on the idea the speech signal constructed based on the source-filter model. Some examples of methods that adopt such model are linear predictive coding (LPC), Mel-frequency Cepstral coefficients (MFCCs), perceptual linear prediction (PLP) coefficients. Although, such method are fine and robust in both speech and speaker recognition. however, the performance decline dramatically with noisy channels since such parameters are very sensitive to changes in speaking conditions [4]. In this paper, the idea is to examine some new parameters that based on the
phase-relying features in terms on speech recognition only. The proposed model used the information presented in phase component in speech signal and encode them as parameters for speech recognition.

As it well-known, the main components presented in speech signal include both the energy parameters and phase parameters [5]. The energy parameter, as it mentioned before, have been used extensively in source-filter model. The phase parameter, on the other hand, paid little attention (neglected) because of the conception that the phase parameter hold or carry no useful information about the speech signal [6]. However, the work presented by [7] suggest that the phase parameter can carry an important information about the speech. However, in the processing of the speech signal this is not accepted idea, since it discard some important features of the speech. Therefore we suggest to use and investigate the phase parameter only since we already knew the importance of using the frequency or the energy parameters on speech signal precessing.

The main objective of this paper is to examine and investigate the potential characteristics of phase-parameter and used them in speech recognition system.

2. Phase parameter model and speech parameterization

In most speech signal processing systems, phase-parameters are usually neglected. The source-filter model, for example, is completely disregard the phase feature of speech when extract encoding parameters of the signal. The reason is the believe that the human perceptual system is deaf to phase. Although this model has led to great advances in speech processing, however, it is known to neglect some important information presented in the speech signal [8]. The phase-model try to compensate the luck of such useful information by examine the effective of using phase parameters in dealing with speech signal processing. Recently, some new work proposed new ways of construction and parameterization the speech signal. These try to take advantage of both spectral and phase features of speech.

An example of these works was proposed by [9] which adopt both energy and phase parameters in terms of text-independent speaker identification under different circumstance. In [10] the authors used the phase information along with the magnitude in order to improve the SNR measure for speech recognition. The work presented in [11] adopted two-dimensional entropy as a feature parameter for speaker identification. In terms of speech recognition, the work proposed by [12] suggest to use the phase space distribution in the time domain of the signal as parameter for speech recognition. Also, the model presented by [13] in to use the phase parameter for speaker identification.

In terms of Arabic language, which is a branch of phonetic speech language that contains combinations of letters, known as phones. A phone represents not continues sound in a language. Phones are used to create phonetic speech that assign how a word should be pronounced or spoken. The study presented in [14] used sub-neural-networks for Arabic speech analysis and shows that the system will achieve better results in the case of the phonologic duration. The work done by [15] is used the neural network in speech and speaker recognition with the MFCC parameters. Also, the method presented by [16] in to use framework the based on the feed-forward neural network in text-independent speaker classification and verification.

In this paper, the model that we suggest is to used the features only presented in the phase parameter and to encode them a signal features for speech recognition. The idea is to explore to what extent that the phase parameter can hold useful information for speech recognition. Generally, in order to estimate the feature coefficients of phase-relaying model, the signal must first pass through a multi-channel filter bank. The filter make it easy for the phase to be estimated around the central frequency of the filter channel. Normally, The outputs of the filter will represent both the spectral and phase coefficient of the signal in that channel. The model will rake in to account only the phase part of the filter to contribute in the feature vector. The reason is to examine the embedded information in the phase parameter. After Passing the speech signal through all the filter channels, the Discrete Cosine transform (DCT) is applied as a final step in the process. The importance of using the DCT stem from the fact that the filter output coefficients are correlate with each other. The DCT is used in this case in order to decorrelate and to eliminate the element of redundancy due to multi channels filter output [17].
The encoding of the input speech signal into a set of feature vectors is done using particular steps. It depends mainly on very general steps in speech signal processing. Usually, it start from dividing the signal into fixed-length frames of 20-30ms each. The frames are passed through a multi-channels complex valued filter bank of central frequencies and bandwidths that vary based on the mel-scale frequency distribution. In our model, these steps that adopted to extract the phase-relaying features of speech signal can be summarized as follows:

(i) Prepare the speech signal, which includes reading the signal and one pole high-pass filter for pre-emphasis high frequencies energy.

(ii) Divide the signal into fixed-length frame and do the winding processing.

(iii) Filter through a bank of $N$ bandpass Gabor filter-bank. Gabor filter is used for optimal representation of the filtered signal in both time and frequency domains.

(iv) For each waveform or one frequency band signal generated by Step 3, calculate the analytic signal using the Hilbert transform.

(v) Calculate the phase of the complex (analytic) signal using:

$$
\phi_n(t) = \arctan\left(\frac{\hat{y}_n(t)}{y_n(t)}\right)
$$

where, the $\phi_n(t)$ represent the phase part of the complex signal $y_n(t)$

The main steps of our suggested model is shown in Figure 1.

After passing the signal frames through all filter channels, the generated feature vector will represent the phase parameter of the signal. It will contain some redundancy values. DCT is used to decorrelate the phase parameter of the signal and produce less redundant coefficients. The DCT equation as described in [18]

$$
Dct(l) = \sum_{n=1}^{N} \phi(n). \cos\left(l \frac{\pi}{N} (n - .5)\right)
$$

where, $M$ is the number of DCT coefficients. In the last stage of processing, the classification stage, a mixture of $M$ Gaussians is used to describe each spoken word. The mixture density is a weighted sum of the $M$ density components Normally, for any input feature vector, the computational and discretion of the conditional probability of Gaussian mixture model is described in [9]:

3. The Speech Database

20 Arabic words pronounced by 10 speakers (5 males and 5 females) is a set of the available data speech are employed. The data is specifically designed and recorded for Arabic tongue speakers that can pronounce correctly with all letters pronounced right. In this corpus, each speaker saying each word 10 times with different times over one month. The speech data is taken in both clean and noisy cases. The test is first done using sound and clear data, then we add some noisy and do the second test under distorted data.

4. Experimental Evaluation and Results

The performance of any model about speech or speaker recognition is undoubtedly based on two main important criteria; first is to choose clear and good database. Second is to choose fine and effective feature extraction model. In our experiment, as mentioned above, the model used phase-parameter of speech to generate feature vector. The evaluation of the model is to use a closed-set of 20 Arabic words selected for speech recognition task. The workshop analysis of our system is to take 3 pronunciation of
Figure 1. Block diagram of the suggested frame-based method.

Each words in the database (the first 3 pronunciations of 3 different speakers) as a training materials. The test materials are extracted from the last 7 pronunciation of the words (7 different speakers).

The parameters of each 3 pronunciations (related to one word in the database) are collected together in one feature vector, which is used later as input to the classified model. The density of the GMM classifier model and the relative weights are estimated later form the training pronunciations.

In our experiments, each spoken word is modelled using 8-GMM components with diagonal covariance matrices. Results are obtained using two constructions of Filterbank (one used 30 channels and other used 40 channels). Also, the utterance is taken in two different style; in clean case and in noisy case.

Fig. 2 shows the speech recognition results obtained using clean and noisy utterance of 30 Channels
Figure 2. 20 isolated words speech recognition accuracy results of clean and noisy database.

Figure 3. clean database speech recognition results of 30 and 40 Gabor filter bank.

Gabor filter with 8 GMM components. The result obtained using modulated speech signal with 25ms framing and hamming windowed signal. only phase- part of the signal contributed in the feature with diagonal Gaussian type is used in classification stage.

Fig. 3 shows a comparison results between 30 and 40 channels filter bank of clean database. As it shown, no huge effect of using different number of bandpass channels in the filter bank. May be the reason implies in the selection of centers frequencies that appear to be similar in both case.

Fig. 4 explain the results of clear data speech with GMM classifier of 5 and 8 GMM components. We can notes the effect of the number of Gaussian components on the results. It is clear that using less GMM components in the classifier will degrade the system performance in all.
Figure 4. Clean database speech recognition results of 30 and 40 Gabor filter bank.

As it appears from the experimental results, we can conclude that ignoring the phase parameter in speech (or speaker) recognition is absolutely wrong. In both cases (speech and speaker recognition) discarding some parts of speech will cause to lose some important information in signal parameterization. This will definitely affect the system performance in total.

5. Conclusion

The work presented in this paper has examined the ability of phase parameters in capturing useful information about speech utterance in terms of speech recognition system. Also, demonstrate that ignoring some parts of speech could cause to lose some crucial information that are critical in system performance. Our experiment shows that the parameterization of the phase components in speech play an important role in estimating the utterance regardless of speaking person. Identifying words or sentences are not easy task, especially when noise involved in the signal or the pronunciation of words is not same or close in the train and test stages. These would put some type of difficulty to identify which speech parameterization are robust to what task we want to do. For this reason, this paper try to adopt and examine part of speech components and evaluate its performance in terms of speech recognition system. Results shows that the phase parameter can contribute greatly in speech recognition. This open the door for improve the selection of the speech features for both speech and speaker recognition systems.

6. Acknowledgements

We would like to thank the Al-Mustansiriyah University/ Baghda-d-Iraq (www.uomustansiriyah.edu.iq) for the support in this work.

References

[1] P. Joseph Campbell, “Speaker recognition: a tutorial,” Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.
[2] V. Mitra, H. Franco, M. Graciarena, and D. Vergyri, “Medium-duration modulation cepstral feature for robust speech recognition,” in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014, vol. 1, pp. 1749–1753.
[3] D. A. Reynolds, “An overview of automatic speaker recognition technology;” in IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-2002, mar 2002, vol. 1, p. 40724075.
[4] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition;” in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 2012, pp. 4117–4120.
[5] Guangji Shi, M. M. Shanechi, and P. Aarabi, “On the importance of phase in human speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1867–1874, Sept 2006.

[6] S. FURUI, *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, New York, USA, 2001.

[7] Tesic Lazar, Bondzulic Boban, Andric Milenko, and Pavlovic Boban, “An experimental study on the phase importance in digital processing of speech signal,” *Acta Polytechnica Hungarica*, vol. 14, no. 8, pp. 197–213, 2017.

[8] Schleusing Olaf, “Multi-parametric source-filter separation of speech and prosodic voice restoration,” 2012.

[9] M. Grimaldi and F. Cummins, “Speaker identification using instantaneous frequencies,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1097–1111, Aug. 2008.

[10] M. S. E. Langarani, H. Veisi, and H. Sameti, “The effect of phase information in speech enhancement and speech recognition,” in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, July 2012, pp. 1446–1447.

[11] B. Bozilovic, B. M. Todorovic, and M. Obradovic, “Textindependent speaker recognition using twodimensional information entropy,” *Journal of ELECTRICAL ENGINEERING*, vol. 66, no. 3, pp. 169173, July 2015.

[12] N. S. Sreekanth, S. N. Pal, G. Arunjith, and N. K. Narayanan, “Phase space point distribution parameter for speech recognition,” in *2009 Third Asia International Conference on Modelling Simulation*, May 2009, pp. 345–347.

[13] V L. Lajish, Kumar R K Sunil, and Pulinholi Vivek, “Speaker identification using a nonlinear speech model and ann,” *International Journal of Advanced Information Technology (IJAIT) 2231-1548*, vol. 2, 11 2012.

[14] S. Selouani and J. Caelen, “Recognition of arabic phonetic features using neural networks and knowledge-based system: a comparative study,” in *IEEE International Joint Symposium on Intelligence and Systems, 1998. Proceedings.*, May 1998, pp. 404–411.

[15] Kaur Gurpreet, Srivastava Mohit, and Kumar Amod, “Speaker and speech recognition using deep neural network,” *International Journal of Emerging Research in Management and Technology*, vol. 6, pp. 118, 06 2018.

[16] Zhenhao Ge, Ananth N. Iyer, Srinath Cheluvanaraja, Ram Sundaram, and Aravind Ganapathiraju, “Neural network based speaker classification and verification systems with enhanced features,” *CoRR*, vol. abs/1702.02289, 2017.

[17] Gandhi Ami and Patil Hemant A., “Novel linear prediction temporal phase based features for speaker recognition,” in *Speech and Computer*, Alexey Karpov, Rodmonga Potapova, and Josif Mporas, Eds., Cham, 2017, pp. 564–571, Springer International Publishing.

[18] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schioler, “Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music,” in *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR)*, 2006.