Evaluation of robustness in ASR for different 'Front-End' methods

Tola S. PhD., Daci A. PhD,
Faculty of Mathematical and Physical Engineering – Polytechnic University of Tirana

saimir_tola@yahoo.com; alfreddaci@gmail.com

Abstract: *Some feature extraction methods suffer performance degradation in different environments. So it has become a necessity to search for new methods that perform better in different types of conditions. Therefore we can make a comparison of the new found methods to evaluate their performance and to determine which is best in multi-condition tests in order to have e more robust ASR system.*

Keywords: FEATURE EXTRACTION, DEGRADATION, PERFORMANCE, ASR.

1. Introduction

Speech recognition, is commonly known as automatic speech recognition (ASR), is the process of converting an acoustic signal, captured by a microphone or a telephone, to a text. The main goal of speech recognition is to get effective ways for mankind to communicate with computers, for example, voice-controlled personal computers. A speech recognizer can be divided mostly in two parts: 'front-end' and 'back-end'. The general structure of a speech recognizer is shown as below.

The purpose of the 'Front-End' is to extract feature vectors from a speech signal. The feature vectors can capture the important characteristics of an utterance. When an unknown utterance is presented a feature vector is obtained. In this paper we study three major feature extractions 'Front-End' in order to get an overview of the situation in which we are studding.

These methods consist in: Gammatone Filter Cepstral Coefficients (GFCC) [1], Mel Frequency Cepstral Coefficient (MFCC) [2], and Perceptual Linear Prediction Coefficients (PLPC) [3]

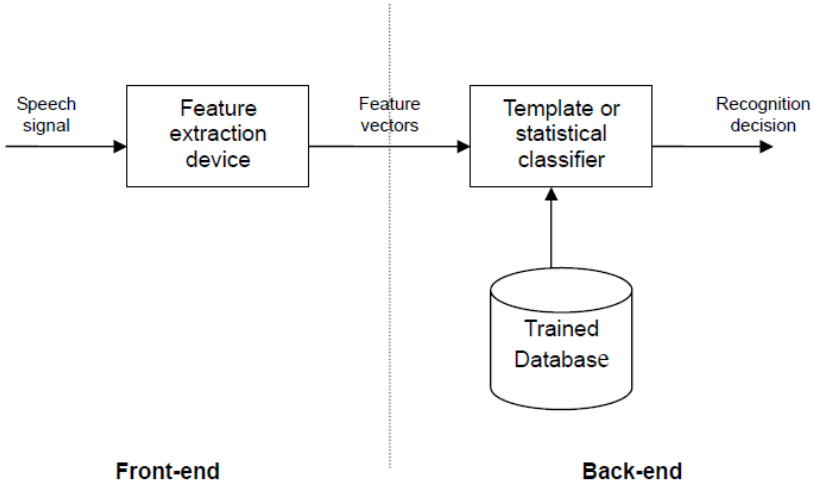

Fig. 1 *Common structure of a speech recognizer*

2. Data Description

One of the major factor that leads to degradations in the performance of ASR systems is the presence of noise in the environment. Such degradations in performance can be due to the mismatch between the conditions in which the systems are trained and the ones in which they are operated.

Some speech enhancement approaches are found really well to deal with unknown noise and filtering such as, Spectral Subtraction, Spectral Normalization. We will see and deal with the parameters that affect the ASR like: pitch, intensity, duration, voice quality, voice strength and the signal to noise ratio [4].

Based on it, data descriptions are shown below:

Type of recognition system: Speaker-independent continuous speech recognition

Front-ends: PLPCC, MFCC, GFCC,

Back-end: Hidden Markov Modelling (HMM) with context-dependent 4-mixture triphone HMMs

Number of coefficients in a vector: 39 (13 static + 13 delta + 13 delta-delta coefficients; Static coefficients include 12 coefficients+ log energy coefficient)

Window size (sec): 20 ms

Step size (sec): 10 ms

Sampling rate: 16kHz

Speech Database: IEEE, Aurora-4

Training set: 4578 utterances spoken by 546 speakers

Test set: 1258 utterances spoken by 134 speakers

Noise Type: White Gaussian noise

SNR range: 0dB, 5dB, 10dB, 15dB, 20dB, clean

3. Experimental Results

To evaluate the performance of the ASR System we will use the WER (Word Error Rate) algorithm. Below we will show the results of our experiment in Table 1 and Fig.2:

Table 1. *The results in percentage of correctly recognized phonemes in various SNR environments*

WER	SNR (dB)					
	0 dB	**5 dB**	**10 dB**	**15 dB**	**20 dB**	**clean**
PLPC	10.72	13.61	18.32	21.44	23.57	62.33
MFCC	12.12	14.97	19.02	23.21	23.57	64.72
GFCC	7.71	10.44	14.21	15.22	17.34	63.31

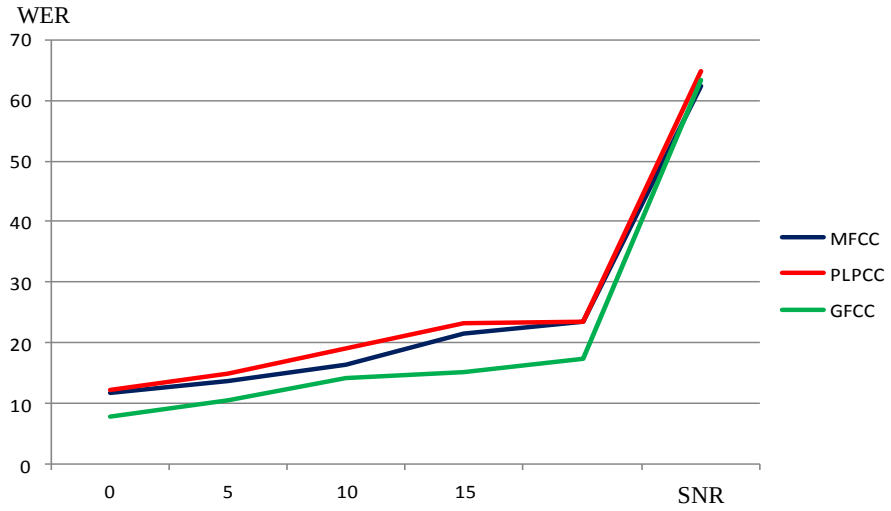

Fig. 2 *Graphical presentation of the results in the in various SNR environments*

4. Conclusions

Amongst the three conventional feature extraction front-ends (MFCC, PLPC and GFCC), it is obvious that PLCC is the worst performing front-end. All of them have up to 62% accuracy in clean environment. Also, the recognition rate of PLPCC in adverse environment is lower than GFCC and MFCC. In noisy conditions (0-20dB SNR), LPCC performs approximately 2-5% worse than PLPCC and MFCC.

Between PLPC and MFCC, MFCC performs slightly better than PLP in general. In all of the above three plots, MFCC performs approximately up to 2% better than PLP, except for the static feature at 5dB SNR, where MFCC is 2% worse than PLP.

GFCC is either performing equally to or better than the three conventional front-ends. In clean, 20dB and 0dB conditions, GFCC has approximately the same recognition rate as MFCC. In all other conditions, GFCC outperforms MFCC by 4-5%.

This method can also be applied in improvement of the measurement uncertainty at reference [5]

5. References

1. Aniruddha Adiga, Mathew Magimai, Chandra Sekhar Seelamantula. Coefficients for Robust Speech Recognition, TENCON conference, pp. 01-04, (2013)
2. Davis, S. and Mermelstein, P., Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. ASSP-28, pp 357-366, (1980).
3. Hermansky, H., Perceptual linear predictive (PLP) analysis for speech, J. Acoust. Soc. Am., pp.1738-1752, (1990).
4. Urmila Shrawankar, Vilas Thakare. The Adverse Conditions and ASR Techniques for Robust Speech User Interface International Journal of Computer Science Issues (IJCSI), 8(5), 440-449 (2011)
5. Klodian Dhoska, Saimir Tola, Agus Pramono, Indrit Vozga, Evaluation of measurement uncertainty for the determination of the mechanical resistance of the brick samples by using uniaxial compressive strength test, Int. J. Metrol. Qual. Eng. 9, 12 (2018)