Arabic Speech Recognition System in Noisy Environments

Hussien A. Elharati[1] Nasser B. Ekreem[2]

Department of Electrical Engineering, Higher Institute of Science and Technology, Sok Aljum'aa, Libya[1]

Department of Electrical & Electronic Engineering, Faculty of Engineering Azzaytuna University – Tarhuna – Libya[2]

الملخص

التعرف على الكلام يعتبرأحد التقنيات الواعدة. تضع اغلب الشركات التكنولوجية العملاقة في العالم الاجهزة التي تدعم الصوت في مقدمة استراتيجيتها. وبالتالي اصبحت هذه التقنية من اكثر مجالات البحث نشاطا. مع ان استخلاص البيانات من الاشارة الصوتية ومعرفة اقرب احتمال للنجاح قد تم تطويره خلال العقد الاخير الا انه لاتزال هناك اشياء كثيرة تحتاج للبحث والتطوير. في ورقتنا هده تم كتابة برنامج الماتلاب لتصميم نظام للتعرف علي الكلام الناطق باللغة العربية بالاعتماد علي استخلاص البيانات بتقنيتي (*Perceptual linear production (PLP) and RASTA-PLP*) **وايجاد تقارب الكلمات والتحقق من النتيجة بتقنية** *Multivariate Hidden Markov Model (HMM)* **واستخدمت في مرحلة التدريب والاختبار النظام بيانات صوتية باللغة العربية اخدت من** 19 **متكلم وكان عدد الكلمات من كل شخص** 24 **كلمة قام المتكلم بنطق الكلمة مرتين وكان العدد الاجمالي للكلمات المسجلة هو** 1368 **كلمة, وفي النهاية باستخدام برنامج الاختبار المعتمد علي تقنية ماركوف كانت النتائج جيدة جدا بنسبة تصل الي** .82.93% **لتقنية** *PLP* **ونسبة تصل الي** *95.06%* **باستخدام تقنية** *RASTA-PLP*

ABSTRACT

Speech recognition can be considered as one of the promising techniques. The world's technology behemoths put voice-enabled devices at the heart of their strategy, and as a result, speech recognition has become one of the most active areas of research. Although feature extraction and likelihood evaluation techniques have been developed and improved over the last decade, they still lack innovation. In this research paper, Multivariate Hidden Markov Model (HMM) investigates the performance of conventional features of perceptual linear production (PLP) and RASTA-PLP to design a robust and reliable Arabic speech recognition system (ASR). For training and testing purposes, the proposed system was evaluated using different noisy data sets of human voice. These small vocabulary isolated speech data set contains the pronunciations of 24 Arabic words. Consonant-Vowel Consonant-Vowel Consonant-Vowel (CVCVCV) structure was recorded from 19 Arabic native speakers, with each speaker saying the same word three times (1368 words). Data is saved separately in a wave file format sampled with 48k sampling rate and 32 bits depth. The system was trained in phonetically rich and balanced Arabic speech words list, 10 speakers * 24 words * 3 times, 720 words total and tested with 9 speakers * 24 words* 3 times *, 648 words total). Using test data word vocabulary, the system obtained a very good word recognition accuracy result of 93.82% using PLP and 95.06% using RASTA-PLP.

Keywords:Feature extraction, likelihood evaluation, speech recognition, Hidden Markov Model, word error rate.

1.INTRODUCTION

Speech Recognition Software ASR distinguishes the speech machine [1]. As illustrated in Figure 1, signal modeling extracts acoustic features from an input

speech signal using a specific algorithm, and then a statistical model matches these features to generate recognition results using classifier techniques [2]. The speech signal is extracted in the front-end into several short frames of 10 to 30 ms length that reflect several useful physical characteristics. The same processes were repeated for all subsequent frames, with a new frame typically overlapping its predecessor by 10 ms to generate a sequence of feature vectors. The back-end selects the most likely words from the trained set and applies statistical modeling to calculate the maximum likelihood. The main challenge involved in designing speech recognition system is to selectthe signal and statistical model, in addition to that, the noise factor considered as a major problem in speech recognition [3]. The goal of this research is to test the effectiveness of noise on features using PLP, RASTA-PLP, and a multi-variate HMM classifier .

The following is how this paper is structured: Section 2: Describe ASR Modules Section 3 discusses the details of the feature extraction techniques PLP and RASTA-PLP, followed by a description of the Hidden Markov Model classifier. Section 4, Conclusions are provided based on a comparison of all of the two methods of speech extraction mentioned above in the previous section.

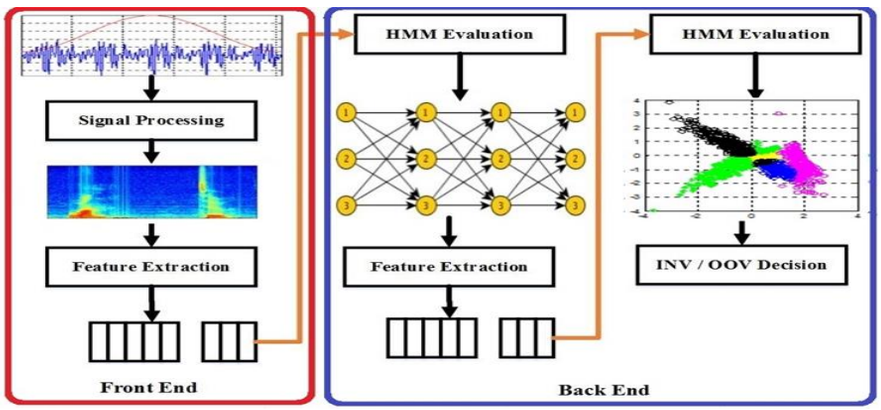


Figure 1 speech recognition system

2. FRONT-END ANALYSIS

To generate the characteristic features of both training and testing data, the acoustic signal was converted into a sequence of acoustic feature vectors using PLP and RASTA-PLP [4].

2.1. Pre-processing

Before extracting the features from the speech signal, several steps were performed, including signal to noise ratio, pre-emphasis, frame blocking, and windowing [5].

2.1.1. Signal-to-Noise Ratio

As shown in Figure 2, the input speech signal has been digitally corrupted by adding various types of realistic noises at SNRs ranging from 30dB to 5dB.

v_addnoise.mMatlab function

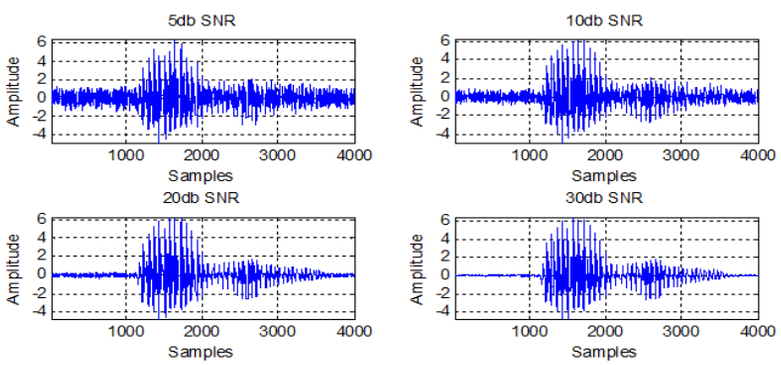


Figure 1. Signal to noise ratio

2.1.2. Pre-emphasis

Pre-emphasis is used to flatten the speech spectrum and compensate for the undesirable high frequency component of the speech signal [6]. The following equation depicts the FIR filter's transfer function.

$$y[n] = x[n] - A\,x[n-1]$$

2.1.3. Blocking and Windowing

Discontinuity the signal at the start and end of each frame by hammering windows typically 25 msec long with a 10 msec shift on pre-emphasized signal y[n], as shown in figure 3.

$w(n) = 0.54 - 0.46\, cos\,(2\pi n/(N-1))$

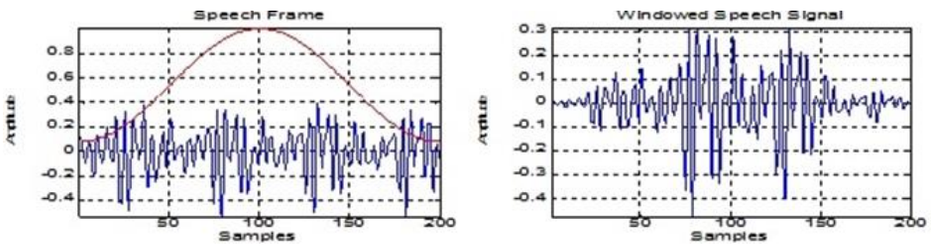


Figure 3. Hamming Window

2.2. Feature extraction

A sequence of feature vectors contains a good representation of the input speech signal by using a feature extraction technique to extract physical characteristics, which is then used to classify and predict new words [7]. Several feature extraction techniques are designed to generate 39 static and dynamic coefficients using Matlab software package.

2.2.1. Perceptual Linear Prediction (PLP)

Based on linear LP analysis of speech, PLP is used to derive a more auditory-like spectrum and calculate several spectral characteristics to match the human auditory system using an autoregressive all-pole model. This type of feature extraction is accomplished by making some assumptions about the psychophysical characteristics of the human hearing process [8].

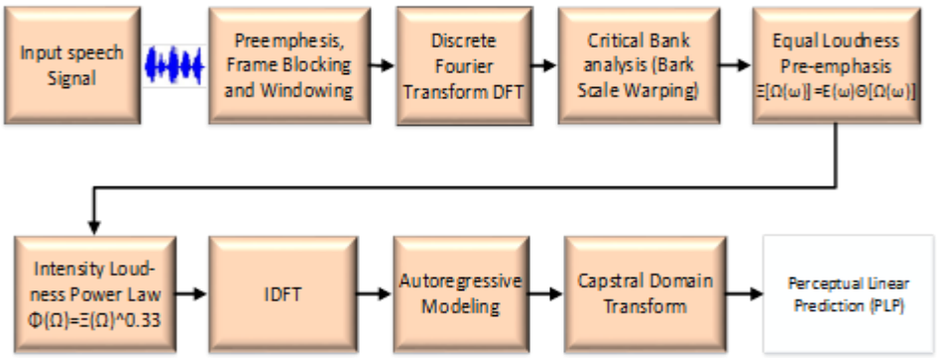
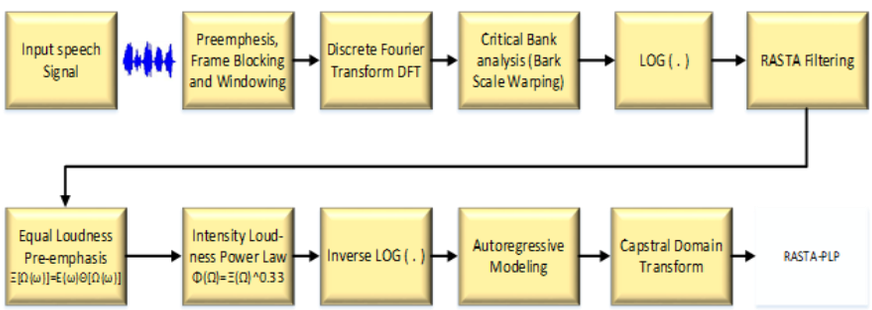


Figure 4. Perceptual Linear Prediction (PLP)

2.2.2. RASTA-PLP

It is an enhanced version of traditional PLP method by filtering the time trajectory in each spectral component [9]. RASTA designed a smooth over short-term noise variations and remove any constant offset in the speech channel, RASTA applies a special band-pass filter to each frequency subband.



3. BACK-END ANALYSIS

3.1. Statistical Modelling process

HMM is used to model non-linearly aligning speech and estimate model parameters [10]. Hidden states Q, observations O, transition probability A, emission probability B, and initial state probability are the parameters of finite-

state machines. HMM was used in this study to classify feature vectors and predict the unknown word.

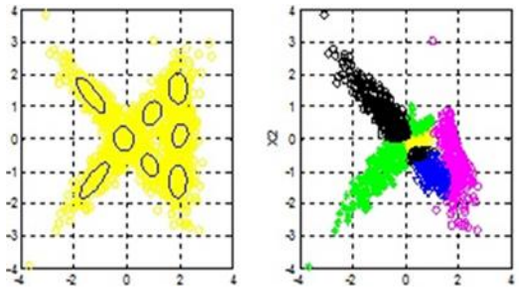


Figure 6. Multi-dimensional Gaussians Hidden Markov Model

3.2. Evaluation process

The probability of the observation sequence was computed using forward-backward dynamic programming, as shown in Figure 6, and the result of the possible state sequence paths was stored in a matrix.

$$P(O \setminus \lambda) = \sum_{i=1}^{N} P(O_t q_t = \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

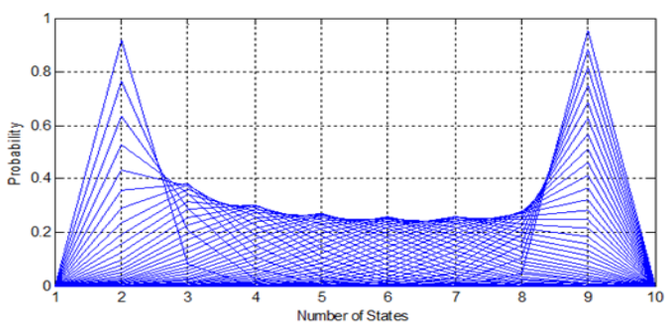


Figure 7. Forward (α) and Backward (β) probability in each state

4. IMPLEMENTATION & RESULTS

In this study, a small vocabulary data set of (24 words * 3 times) Arabic CVCVCV words is recorded from 19 adult male speakers (total 1368) then divided into training and testing files, as illustrated in table 1.

Table 1. CVCVCV Arabic words

No.	Word	No.	Word	No.	Word	No.	Word
1	فَعَلَ	7	فَعِلَ	13	فَعُلَ	19	فُعِلَ
2	رَفَعَ	8	بَخِلَ	14	بَلُغَ	20	ذُكِرَ
3	ذَكَرَ	9	عَمِلَ	15	صَلُحَ	21	جُمِعَ
4	ذَهَبَ	10	حَفِظَ	16	سَهُلَ	22	خُلِقَ
5	شَرَحَ	11	سَمِعَ	17	كَبُرَ	23	كُتِبَ
6	كَتَبَ	12	فَرِحَ	18	كَرُمَ	24	حُشِرَ

The proposed speech recognition model's performance was evaluated. PLP and RASTA-PLP conventional feature extractions were trained and tested in noisy conditions to find the best word recognition rate using a Multivariate Hidden Markov Model (HMM) classifier. Several experiments are conducted in various conditions with small vocabulary isolated words corpora (19 people *24 words * 3 times). Table-2 shows the resulting confidence level intervals for the obtained recognition rate. All data were trained with four to ten states and modeled with an eight-dimensional Gaussian Hidden Markov Model. Figure 7 depicts the recognition rate obtained for each feature extraction method, whereas figure 8 reveals the overall recognition rate using PLP and RASTA-PLP features percentage.

Table 2. Overall recognition rate using PLP features

State No.	Total error count	Total correct count	Recognition rate
4	130	518	79.93
5	90	558	86.11
6	82	566	87.34
7	80	568	87.65
8	48	610	94.13
9	44	604	93.20
10	40	608	93.82

Table 3. Overall recognition rate using RASTA-PLP features

State No.	Total error count	Total correct count	Recognition rate
4	126	522	80.55
5	86	562	86.72
6	78	570	87.96
7	71	577	89.04
8	40	608	93.82
9	36	612	94.44
10	32	616	95.06

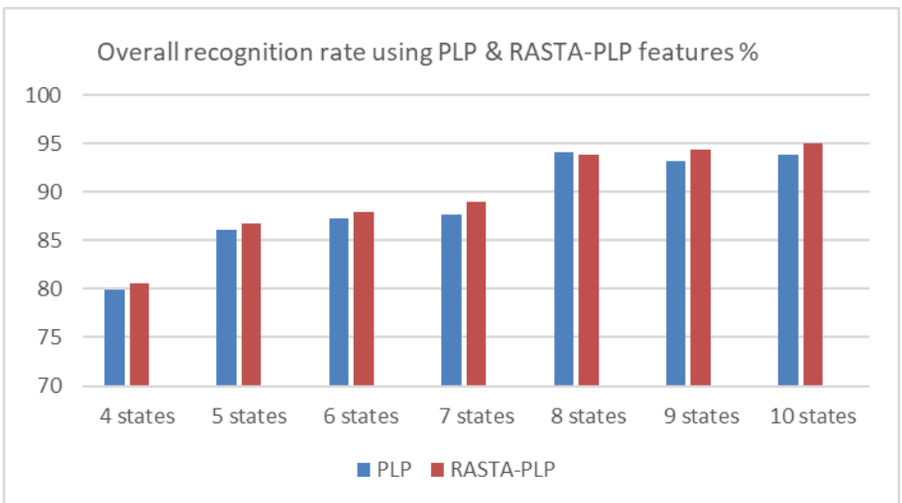


Fig. 8. Overall recognition rate using PLP & RASTA-PLP features

5. CONCLUSION

The goal of this study is to compare the performance of two feature extraction techniques, PLP and RASTA-PLP, using a discrete-observation HMM-based isolated word recognizer, implemented using MATLAB software package. After dividing the audio signal into 4 to 10 different states and clustered by 8 Gaussian mixtures, the results show that the acoustic signals extracted using RASTA-PLP granted the best recognition rate at 10 states and 8 Gaussian mixtures

(95.06%), whereas PLP provided the highest recognition rate at 10 states and 8 Gaussian mixtures. According to several experiments, results show that RASTA-PLP produced the best results because it applies a special band-pass filter to each frequency sub-band to smooth over short-term noise variations and eliminating any constant offset in the speech channel.

6. REFERENCES

[1] Tambe, T., Yang, E. Y., Ko, G. G., Chai, Y., Hooper, C., Donato, M., ... & Wei, G. Y. (2021, February). 9.8 a 25mm 2 soc for iot devices with 18ms noise-robust speech-to-text latency via bayesian speech denoising and attention-based sequence-to-sequence dnn speech recognition in 16nm finfet. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)* (Vol. 64, pp. 158-160). IEEE.

[2] Hussien, A. E., Mohamed, A., & Veton, Z. K. (2020). Arabic Speech Recognition System Based on MFCC and HMMs. *Journal of Computer and Communications*, *8*(03), 28-34.

[3] Vos, T. G., Dillon, M. T., Buss, E., Rooth, M. A., Bucker, A. L., Dillon, S., ... & Dedmon, M. M. (2021). Influence of protective face coverings on the speech recognition of cochlear implant patients. *The Laryngoscope*, *131*(6), E2038-E2043.

[4] Këpuska, V. Z., & Elharati, H. A. (2015). Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions. *Journal of Computer and Communications*, *3*(06), 1.

[5] Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, *87*(4), 1738-1752.

[6] T. Sainath, A. rahman Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep Convolutional Neural Networks for LVCSR,” in ICASSP, 2013.

[7] Këpuska, V. Z., & Elharati, H. A. (2015). Performance Evaluation of Conventional and Hybrid Feature Extractions Using Multivariate HMM Classifier. *International Journal of Engineering Research and Applications*, *5*(4), 96-101.

[8] Chiu, C. C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., ... & Bacchiani, M. (2018, April). State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 4774-4778). IEEE.

[9] Liu, B., Cai, H., Zhang, Z., Ding, X., Wang, Z., Gong, Y., ... & Yang, J. (2021). More is less: Domain-specific speech recognition microprocessor using one-dimensional convolutional recurrent neural network. IEEE Transactions on Circuits and Systems I: Regular Papers, 69(4), 1571-1582.

[10] Elharati, H. (2019). *Performance evaluation of speech recognition system using conventional and hybrid features and hidden Markov model classifier* (Doctoral dissertation, Florida Institute of Technology).

[11] Green, J. R., MacDonald, R. L., Jiang, P. P., Cattiau, J., Heywood, R., Cave, R., ... & Tomanek, K. (2021). Automatic Speech Recognition of Disordered Speech: Personalized Models Outperforming Human Listeners on Short Phrases. In Interspeech (pp. 4778-4782).

[12] Canfarotta, M. W., Dillon, M. T., Buchman, C. A., Buss, E., O'Connell, B. P., Rooth, M. A., ... & Brown, K. D. (2021). Long-term influence of electrode array length on speech recognition in cochlear implant users. The Laryngoscope, 131(4), 892-897.

[13] Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech, and signal processing (pp. 6645-6649). Ieee.

[14] Kim, C., Misra, A., Chin, K., Hughes, T., Narayanan, A., Sainath, T., & Bacchiani, M. (2017). Generation of large-scale simulated utterances in virtual

rooms to train deep-neural networks for far-field speech recognition in Google Home.

[15] Huggins-Daines, D., Kumar, M., Chan, A., Black, A. W., Ravishankar, M., & Rudnicky, A. I. (2006, May). Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In 2006 IEEE international conference on acoustics speech and signal processing proceedings (Vol. 1, pp. I-I). IEEE.

[16] Murr, A. T., Canfarotta, M. W., O'Connell, B. P., Buss, E., King, E. R., Bucker, A. L., ... & Dillon, M. T. (2021). Speech recognition as a function of age and listening experience in adult cochlear implant users. The Laryngoscope, 131(9), 2106-2111.

[17] Toscano, J. C., & Toscano, C. M. (2021). Effects of face masks on speech recognition in multi-talker babble noise. PloS one, 16(2), e0246842.

[18] Elharati, H. (2019). Performance evaluation of speech recognition system using conventional and hybrid features and hidden Markov model classifier (Doctoral dissertation, Florida Institute of Technology).

[19] Elharati, H. A. Marghani A. K, Elhaj A. M., sims A. J. (2015). Fractal analysis on the detection of the malignancy changes of pancreatic cancer. *The Libyan Journal of Science*, *26*(1), 24-32.