Performance Evaluation of Hybrid Features in ASR System

Hussien A Elharati
Department of Electrical Engineering, Higher Institute of Science and Technology, Sok Aljum'aa, Libya

Helharati2013@my.fit.edu

Nasser B. Ekreem
Faculty of Engineering, Department of Electrical & Electronic Engineering, Azzaytuna University - Tarhuna – Libya
nekreem@gmail.com

Khaled A Marghani
Department of Physics, Faculty of Science, University of Tripoli, Libya

K.marghani@uot.edu

Ahmed Ali Alsoukni
Department of Electrical Engineering, Higher Institute of Science and Technology, Sok Aljum'aa, Libya
Ahmedalsoukni@gmail.com

الملخص

عملية استخلاص الميزات وتصنيفها وتقييمها من الاهتمامات الرئيسية في نظام التعرف على الكلام وما تزال تعتبر من أكثر مجالات البحث نشاطًا في وقتنا الحاضر. في هذا البحث تم فحص أداء خوارزمية جديدة لنظام هجين لاستخلاص خصائص الصوت استخدمت فيه عدد اربع تقنيات هي Prediction Cepstrum Coefficient, perceptual linear production, Mel Frequency Cepstrum Coefficient, RASTA–PLP. تم تصنيف البيانات المستخلصة باستخدام هذه التقنية الهجينة وتقييمها باستخدام تقنية Hidden Markov Model (HMM) واستخدم لتقييم أداء هذه الخوارزمية المقترحة مجموعة كبيرة من البيانات الصوتية تتكون من إحدى عشرة كلمة باللغة الانجليزية (صفر إلى تسعة زائد الحرف أو) والتي تم تسجيلها من عدد 4558 متحدثًا بالغًا، سجلت كل كلمة من كل شخص مستهدف مرتين وكان زمن اخد العينة Sampling Rate هو 8 كيلو هرتز وتم حفظها في 4558 ملف منفصل امتداده WAVوقسمت كل هذه الاصوات الي مجموعتين مجموعة للتدريب ومجموعة للتختبار

وقد اعطت التقنية الهجينة MFCC+RASTA أفضل نتيجة بنسبة مقدارها 99.43 وبعدد اخطاء يساوى 14 خطأ من أصل 2472 صوت في العينة.

الكلمات الدليلية: استخراج الميزة ، وتقييم الاحتمالية ، والتعرف على الكلام ، ونموذج ماركوف المخفي ، ومعدل الخطأ في الكلمات

ABSTRACT

Feature extraction, classification, and evaluation processes are considered the main concerns in speech recognition system and still the most active area of research nowadays. In this paper the performance of new hybrid feature extraction algorithm is examined using Linear Prediction Cepstrum Coefficient (LPCC), perceptual linear production (PLP), Mel Frequency Cepstrum Coefficient (MFCC), and RASTA-PLP. The extracted data vectors are classified and evaluated using Hidden Markov Model (HMM). The performance of the proposed hybrid algorithm is assessed using data set of human voice, which consists from eleven words (zero to nine plus oh) and recorded from 4558 adult speakers, each person said the same word 2 times. The collected data are sampled by 8 KHz and saved in 4558 WAV files divided into training and testing data. According to the final results, the proposed system provided an excellent recognition rate with 99.43% using the combination between MFCC and RASTA.

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

1. INTRODUCTION

Due to the advances in communication systems, machines are able to identify and interact with human instructions [1]. In general the Speech Recognition system usually divided into two parts, Front-End and Back-End as illustrated in Fig. 1.

Front-end was designed to generate the acoustic features from speech signal using specific algorithms, whereas Back-End matches these features to achieve and generate the recognition result [2]. In our research, the speech signal divided into number of frames 25ms each and overlaps with 10ms in order to generate the feature vectors and then stored in specific matrices. At the other hand, Back-End sorts those vectors and calculates the maximum likelihood to select the most likely sequences of the phonemes [3].

In this work, the performance of a new hybrid feature extraction algorithm is required. It is more beneficial to investigate the performance of hybrid Front-End features. HMMs with Gaussian mixture are used as a classifier through this research. This work is organized into 6 sections; section 2 offers a background on front end proposed analysis. In section 3, Back-end analysis and the development of HMM classifier is provided. Section 4 is devoted to discuss the results and analysis of the data. Eventually, conclusion and references are presented in sections 5 and 6 respectively.

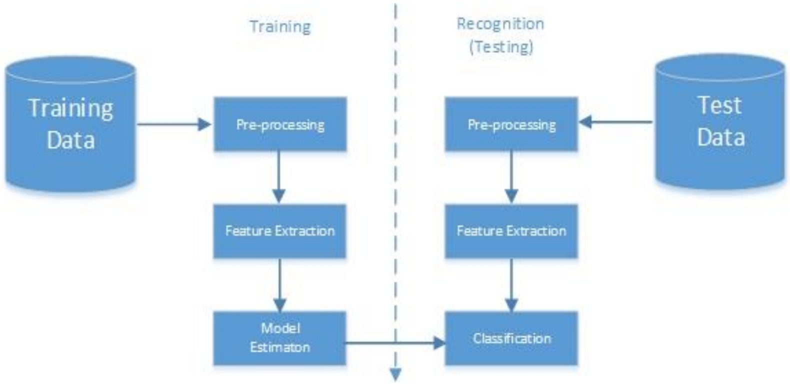

Figure 1: Front-End and Back-End in ASR

2. FRONT-END PROPOSED ANALYSIS

Front-End analysis controls the process of converting speech acoustic signal into a sequence of acoustic feature vectors. The conventional feature extractions methods, MFCC, LPCC, PLP, and RASTA-PLP are used for the evaluation of the performance of proposed features. In order to be ready for feature extraction calculations, pre-emphasis, frame blocking and windowing were taken onto speech signal [4].

Using high pass FIR filter Signal was Pre-emphasized as shown in Equation (1) to flatten speech spectrum and compensates the unwanted high frequency part of the speech signal.

$$Y[n] = x[n] - A\,x[n-1] \qquad (1)$$

Where,

$x[n]$: Input signal

$x[n-1]$: Previous speech signal

A: Pre-emphasis factor = 0.975.

Frame Blocking and Windowing are used to reduce signal discontinuities at the beginning and the end of the extracted frames. Hamming window function typically 25 ms long and 10 ms shift is selected and applied on the speech signal in order to cover and grape the data using equation (2) as depicted in Fig. 2.

$$w(n) = 0.54 - 0.46\,cos\,(2\pi n\,/\,(N-1))\; 0 \leq n \leq N-1 \qquad (2)$$

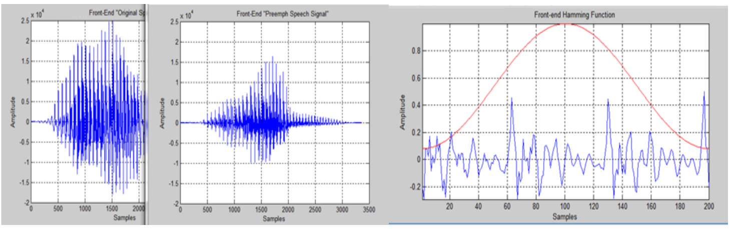

Figure 2: Original signal, pre-emphasis, and Hamming Window

2.1 Conventional features

It is a sequence of feature vectors that carries a good representation of the speech signal [5]. In this work a new hybrid features were extracted using Matlab software package. Each extraction method generated 39 parameter coefficients in one vector, 12 static parameters, 1 power parameter, and 26 dynamic parameters (13 first derivative, and 13 second derivative).

To generate the features using MFCC algorithm, spectral features were extracted from frames sequence using Fast Fourier Transform in order to obtain 256 point certain parameters from each frame and then convert the power-spectrum to a Mel-frequency spectrum, and take the logarithm of that spectrum and compute its inverse Fourier transform.

Based on linear LP analysis, PLP algorithm is used to develop more auditory-like spectrum, and calculate several spectral characteristics in order to match human auditory system using autoregressive all-pole model.

At low bit-rate, LPC works to represent an attempt to mimic human speech using auto-correlation method and Levinson-Durbin recursion. To generate the features, LPC were calculated by approximating the current sample as a linear of the past sample as represented in Equation (3), and then converting the LPC parameter into cepstral coefficients [7].

$$R(i) \;=\; \sum N\; n = w1 - 1\; sw(n)\; sw(n \;-\; i)\; 0 \;\le\; i \;\le\; p \qquad (3)$$

Where;

N_w, represents the window length

s_w, denotes windowed segment.

By filtering the time trajectory in each spectral component, RASTA-PLP is accomplished. RASTA-PLP technique considered as more advanced compared with traditional PLP. A special band pass filter as shown in Equation (4) was applied to each frequency subband in order to smooth-over short-term noise variations and eliminates any constant offset in the speech channel [8].

$$H(z) = 0.1 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98z^{-1)}} \qquad (4)$$

2.2 Hybrid features

The first 20 features from the MFCC, LPCC, PLP and RASTA-PLP mattress are taken to collect and generate 40 hybrid features. Each feature produces 20 parameter coefficients; each two different feature methods are grouped in one vector in order to provide 40 parameter coefficients as shown.

- LPCC and PLP.
- MFCC LPCC.
- MFCC and RASTA-PLP.
- MFCC and PLP.

3. BACK-END ANALYSIS

After extracting the relevant characteristics from the speech signal and stored in vectors, the powerful statistical tool, the Hidden Markov Model (HMM) is used to classify these features in order to find the correct recognition decision. In our research, HMM is used because of its ability to model non-linearly aligning speech and estimate the model parameters [9]. Also, in this work, 10 Gaussian mixtures were used to model the emission probability distribution in each hidden state. In training process, the transition probability matrix, observation parameters, prior probabilities, and Gaussian distribution were re-estimated to find the new parameters which were used to generate the likelihood scores. These scores are used to find the best bath between frames to recognize the unknown word.

In our HMM design, the first concern is the evaluation of the probability that any sequence of states has produced the sequence of observations. Using Equation (5), forward (α) and backward (β) algorithms were used to determine the overall result of the possible state sequence paths.

$$P(O\backslash\lambda) = \sum_{i=1}^{N} P(O_t q_t = \lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \qquad (5)$$

The mean, covariance of transition probability matrix, and Gaussian mixture parameters were re-estimated using Baum-Welch algorithm. Multi-dimensional Gaussian PDF is achieved using Equation (6).

$$p(x\backslash\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}\Sigma^{1/2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \qquad (6)$$

Where:
d: Number of dimensions
x: Input vectors
μ: Mean vector
Σ: Covariance matrix.

Baum welch algorithm, as represented in Equation (7) is the mathematical model which was used to learn and encode the current observation sequence to recognize similar and new observation sequence in training process.

$$\lambda^{*}=\arg\max_{\lambda}[P(O\,|\,\lambda)] \qquad (7)$$

Viterbi algorithm also applied in order to determine the optimal scoring path of the state sequence in decoding process [10]. Equation (8) defines maximal probability of state sequences. On the other hand, Equation (9) was used to obtain the optimal scoring path of state sequence between frames used 8- states HMM as shown in figure 3.

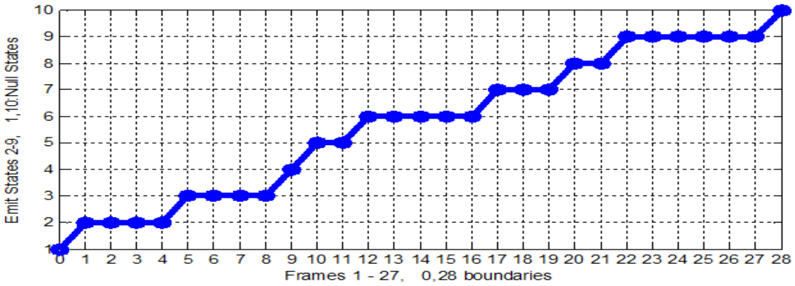

Figure. 3 Viterbi trellis computation for 8-states HMM

$$\delta\ t(i)\ =\ max(P(q(1),q(2),..,q(t-1);o(1),o(2),..,o(t)|\lambda) \qquad (8)$$

$$q*_T=\arg\max_{1\le i\le N}[\delta_T(i)] \qquad (9)$$

4. RESULTS AND CONCLUSION

In this work, the performance evaluation of hybrid feature extraction techniques of MFCC, LPCC, PLP, RASTA-PLP is achieved, and in return, maximum word recognition rate using Multivariate Hidden Markov Model (HMM) classifier was obtained.

Small vocabulary isolated words corpora were used in the experiments, which consists of eleven words (zero to nine and oh) that were recorded from male and female adult speakers. Hybrid features were extracted and used to obtain results of confusion matrix of the average classification. The extracted features were trained and tested using 12 states and modeled by 3-10 multi-dimensional Gaussians Hidden Markov Model as tabulated in Table 1, which shows the summary of the recognition rate obtained from each hybrid feature extraction method experiment.

Table 1. Recognition rate of Hybrid feature extractions

Feature Extraction method	Total error count	Total correct count	Recognition rate %
LPCC+PLP	18	2468	99.15
MFCC+LPCC	21	2465	99.27
MFCC+RASTA	14	2472	99.43
MFCC+PLP	18	2468	99.15

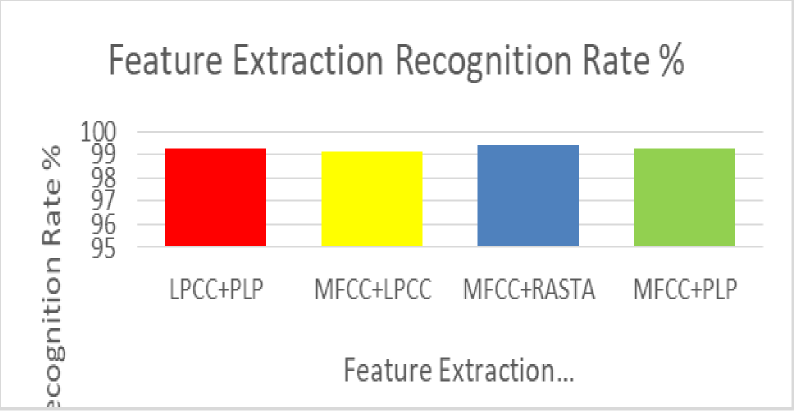

Figure. 4 Overall recognition rate of Hybrid feature extraction

The performance evaluation of the hybrid features were carried out by implementing a discrete-observation multivariate HMM-based isolated word recognizer in MATLAB using 4558 isolated recorded speech data, divided into 2072 training and 2486 testing data set.

With respect to Fig. 4, results emphasized that acoustic signals that were extracted using the hybrid algorithms of (MFCC, LPCC, PLP, and RASTA-PLP) provided excellent recognition rate using 12 states and 4 Gaussian mixtures. The combination of MFCC and RASTA provided 99.43% recognition rate, which considered being the best recognition rate among other features sorted by 12 states and 4 Gaussian mixtures. On the other hand, hybrid combination of MFCC and LPCC provided 99.27 recognition rate using the same states and Gaussian mixtures numbers. Finally the Combination of LPCC, PLP and the combination of MFCC and PLP represented the third highest recognition rate at 99.15% using 10 states and 3 Gaussian mixtures, and this was the lowest o recognition rate achieved.

5. REFERENCES

[1] Elharati, H. ,Alshaari, M. and Këpuska, V. (2020) Arabic Speech Recognition System Based on MFCC and HMMs. Journal of Computer and Communications, 8, 28-34. doi: 10.4236/jcc.2020.83003.

[2] Këpuska, V.Z. and 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, 1-9. https://doi.org/10.4236/jcc.2015.36001

[3] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, Jan 2012.

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

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

[6] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition,” in ICASSP, 2016.

[7] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-End Attention-based Large Vocabulary Speech Recognition,” in ICASSP, 2016.

[8] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, N. Jaitly, B. Li, J. Chorowski, and M. Bacchiani, “State-of-the-art Speech Recognition With Sequence-to-Sequence Models,” in ICASSP, 2018.

[9] N. Kanda, R. Takeda, and Y. Obuchi, “Elastic spectral distortion for low resource speech recognition with deep neural networks,”in ASRU, 2013.

[10] A. Ragni, K. M. Knill, S. P. Rath, and M. J. F. Gales, “Data augmentation for low resource languages,” in INTERSPEECH, 2014.

[11] N. Jaitly and G. Hinton, “Vocal Tract Length Perturbation (VTLP) improves speech recognition,” in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.

[12] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, “Deep Speech: Scaling up end-to-end speech recognition,”in arXiv, 2014.

[13] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, “Audio Augmentation for Speech Recognition,” in INTERSPEECH, 2015.

[14] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, “Generation of large-scale

simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in INTERSPEECH, 2017.

[15] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks,"in ICASSP, 2015.

[16] Elharati, H.: Performance evaluation of speech recognition system using conventional and hybrid features and hidden Markov model classifier. PhD Thesis, College of Engineering and Science of Florida Institute of Technology (2019)

[17] Alshaari, Mohamed, HussienElHarati, and VetonKepuska. Phonemes of Arabic LDC2020S13. Web Download. Philadelphia: Linguistic Data Consortium, 2020.