Application of Gaussian Mixture Models for feature extraction

Brotoń Grzegorz, Tobiasz Maciej
AGH University of Science and Technology

***Abstract*—Feature extraction is a very important process which is mandatory for almost every complex data analysis, which human speech definitely is. This paper describes how Gaussian Model Mixture can be used for the feature extraction from audio samples representing isolated words. GMM features are compared with MFCC method, which is the most popular one for this problem. Speech recognition systems are not a novelty, but as a result of outstanding results they achieve, they are used in increasingly broader scope of applications. Therefore research in this area is still needed, looking for potential improvements.**

***Index Terms*—Gaussian Mixture Models, MFCC, feature extraction, speech recognition**

I. Introduction

Speech recognition or automatic speech recognition (ASR) systems attempt to map from a speech signal to the corresponding sequence of words it represents using pattern recognition algorithms. To make this possible, a series of acoustic features are extracted from the speech signal first. This is typically performed in *pre-processing* phase, which one of the roles is to prepare data for further analysis. Transformation of the data into some new space of variables is called *feature extraction*. This process greatly reduces the variability within data, because the scale and a size of all samples are now the same, which makes it much easier for a subsequent pattern recognition algorithm to distinguish between the different classes. Pre-processing might also be performed in order to speed up computation, which is crucial for real-time system, like Real-Time ASR (RASR).

The structure of this paper is as follows. Section II outlines current standards and techniques related to speech recognition systems and feature extraction for speech signals. Section III briefly presents a theory behind the human speech and MFCC features extraction. The Gaussian Mixture Models and EM algorithm are described in section IV. Section VI presents results of the work and conclusions.

II. Related work

Over the years a number of different methodologies have been proposed for isolated word and continuous speech recognition. First published researches were made many years ago, in times when currently known technology and computers was in its infancy. Example is work [1] published in 1978, which presents a system for speech recognition based on linear predictions. Since then, many different methods were developed and tested in speech recognition domain. Taking under consideration more contemporary publications, [2] presents a brief comparison of different methods for feature extraction used for speech and sound signals. Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT) and Perceptual Linear Prediction (PLP) methods are discussed in this paper. [3] describes an approach of speech recognition using the features extracted using MFCC from speech signal of spoken words. Principal Component Analysis (PCA) is employed as the supplement in feature dimensional reduction state prior to training and testing speech samples via Maximum Likelihood Classifier (ML) and Support Vector Machine (SVM). In [4] a practical application of MFCC feature extraction for speaker recognition system from very small length of audio signals is shown. Despite the final efficiency at level 99.5%, the author points out that the performance of MFCC is affected by the number of filters, the shape of filters, the way that filters are spaced, and the way that the power spectrum is warped. In [5] authors present usage of features extracted using MFCC as a voice signal representation. Several methods such as Liner Predictive Predictive Coding (LPC), Hidden Markov Model (HMM), Artificial Neural Network (ANN) evaluation are presented. [6] presents usage of Mixtures of Gaussians for formant analysis. In this work, the formant parameters are represented in the form of Gaussian mixture distributions, estimated from the Discrete Fourier Transform (DFT) magnitude spectrum of the speech signal. In [7] usage of Gaussian Mixture Model for feature extraction for speech recognition is shown. Author presents various forms of GMM feature extraction. Results are presented in comparison to the state of the art MFCC and PLP methods. What is more, system using combination of GMM and MFCC features is presented.

III. Speech recognition and feature extraction

A. Human speech

Human speech is the result of the manipulation of human vocal folds in cycles of opening them and closing. [8] In the result, waves of frequency 125 Hz for men and 210 Hz for women are generated. This fundamental frequency impacts the perceived *pitch* of the voice. Human speech can be classified as voiced and voiceless sounds, where both of them produce sounds. The difference between them is that with the first one additional vibrations of the speech organs are related. The key component in speaking is the vocal tract which composed of the oral and the nasal part, which acts as a resonator. Both voiced and voiceless sound are further modulated by articulation and create different resonances by the vocal tract. For pronunciation, we split a word into *syllables*. [9] A syllable usually contains one vowel sound, with or without surrounding consonants. Consonants are sounds that are articulated with a complete or partial closure of the vocal tract. It can be voiced or voiceless. One the other hand, vowels are syllabic speech sounds that are pronounced without any obstruction

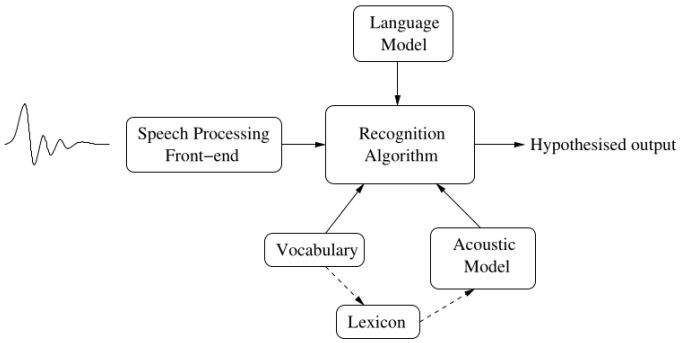

Fig. 1. General speech recognition system

in the vocal tract. The same letter in different words can be pronounced differently. A *phoneme* is a unit of sound that distinguishes one word from another in a particular language. For example, American English has about 44 phonemes. *Phones* are the acoustic realization of phonemes. *Allophones* are a kind of phoneme that changes its sound based on how a word is spelled. There can be many allophones for the same phonemes. Phonemes are an abstract concept in linguistic to distinguish words, while phones are how human pronounce them. [10] For audio signals analysis frequency domain is preferred. When analyzing speech signals, a few dominant frequencies can be observed, which are called *formants*. [11]

B. Principle of speech recognition

The main objective of a speech recognition system is finding the best sequence of words corresponding to the audio based on the acoustic and language model. Statistical pattern recognition is the current paradigm for automatic speech recognition.

$$\mathbf{w}^* = \underset{w}{\operatorname{argmax}}\ P(\mathbf{w}|\mathbf{x}) = \underset{w}{\operatorname{argmax}}\ P(\mathbf{x}|\mathbf{w})P(\mathbf{w}) \qquad (1)$$

where:

$\mathbf{w} = w_1, w_2, \ldots, w_m$- word sequence
$\mathbf{x} = x_1, x_2, \ldots, x_n$ - acoustic observations

In 1 $P(\boldsymbol{x}|\boldsymbol{w})$ is called acoustic model and $P(\boldsymbol{w})$ can be interpreted as a language model. To create an acoustic model, observation **x** is represented by a sequence of acoustic feature vectors (x_1, x_2, x_3, ...) obtained in feature extraction process. An overview of a speech recognition system is given in figure 1. [7]

The most common approach to the problem of classifying speech signals is the use of *hidden Markov models* (HMMs). One advantage of using HMMs is that they are a statistical approach to pattern recognition. This allows a number of techniques for adapting and extending the models. Furthermore, efficient recognition algorithms have been developed for HMM based approach. [12] Various algorithms based on *neural network* (NN) ideas as alternatives to HMM have been proposed in [13], however HMM based system will be presented in this paper.

C. MFCC feature extraction

The role of feature extraction is to remove redundant information from voice signal, make extracted features independent and capture the dynamics of phones – the context. [14] One of the most popular audio feature extraction method is the Mel-frequency Cepstral Coefficients (MFCC) which generates vectors of 39 features. [5]

The whole process is preceded by some pre-processing. Audio signal is a continuous analog signal and therefore it has to be converted into discrete one in process called *sampling*. For speech recognition a sampling of frequency of 8 kHz or 16 kHz is used. [15] It is followed by a *pre-emphasis* process, which main role is to boost the amount of the energy in the high frequencies. For voiced segments, there is more energy at the lower frequencies than the higher frequencies. This is related to the glottal source and is called spectral tilt. [16] Next, audio clip is divided into frames using a sliding window technique. There are generated windows 25 ms long and advanced every 10 ms. However input signal cannot be cut like this, because the suddenly fallen in amplitude will create a lot of noise that shows up in the high-frequency. To prevent this, a Hamming or Hanning windows in the time domain are applied. Windowing process is performed, because context is very important in speech. Pronunciations are changed according to the articulation before and after a phone. [7] The last step is to convert the signal from time domain into the frequency domain. Therefore, the Discrete Fourier Transform (DFT) is applied to the input signal. [17]

First 12 of the features are related to the amplitude of frequencies (formants). They are generated based on *Mel scale* model. The equipment measurements are not the same as human hearing perception. The perceived loudness changes according to frequency and perceived frequency resolution decreases as frequency increases – humans are less sensitive to higher frequencies. Triangular band-pass filters coverts the frequency information to mimic what humans would perceive. In this model, all mappings are non-linear. Mel frequency model is applied to the pre-processed DFT power spectrum signal in process called *Mel Binning*. The output for each Mel-scale power spectrum slot represents the energy from a number of frequency bands that it covers. [18]

$$\mathbf{y}_m = \sum_{k=0}^{N-1} \mathbf{w}_k |\mathbf{x}_k|^2 \qquad (2)$$

where:

k - DFT bin number (0,...,*N-1*)
m- mel-filter bank number (0,...,*M-1*)
N- number of triangular mel weighting filters
w - weight given to the *m*th output band

In the next step, the log of the Mel filterbank output is taken to imitate human hearing organs perception. Computation of the Cepstral is performed to separate information related to phones and the pitch. Pitch varies with people, but it is has not much impact for speech recognition. It is the first observed

formant, called *F0*. Due to its little information for speech recognition, it should be removed. Discrete Cosine Transform (DCT) applied to the transformed mel frequency coefficients produces a set of cepstral coefficients. Prior to computing DCT the mel spectrum is usually represented on a log scale. This results in a signal in the cepstral domain with a que-frequency peak corresponding to the pitch of the signal and a number of formants representing low que-frequency peaks. Most of the signal information is represented by the first few MFCC coefficients, therefore higher order DCT components can be ignored. DCT is an orthogonal transformation, therefore the transformation produces uncorrelated features. Finally, formula for MFCC can be written as 3.

$$\mathbf{y}_n = \sum_{m=0}^{M-1} log(\mathbf{y}_m) cos(\frac{\pi n(m-\frac{1}{2})}{M}) \tag{3}$$

where:

n = 1,2,...,C; C - number of features

There can be a different number of features extracted from the results of the DCT. Traditional MFCC systems use only 8–13 cepstral coefficients, in this paper we present it as 12.

The 13th feature is the energy of each frame. The energy in a frame for a signal **x** in a window from time sample t_1 to time sample t_2, is represented at the equation below:

$$E = \sum_{t=t_1}^{t_2} \mathbf{x}_t{}^2 \tag{4}$$

In pronunciation, context and dynamic information are important. Characterizing feature changes over time provides the context information for a phone. Another 13 values compute the delta values **d** below. Delta coefficients tell about the speech rate.

$$\mathbf{d}_t = \frac{\mathbf{c}_{t+1} - \mathbf{c}_{t-1}}{2} \tag{5}$$

The last 13 parameters are the dynamic changes of **d** from the last frame to the next frame. They are also called delta-delta coefficients and can be interpreted as a second-order derivative of **c**. Delta–delta coefficients provide information similar to acceleration of speech. [19]

IV. Gaussian Mixture Models

A. GMMs in general

Gaussian Mixture Model (GMM) is a probabilistic model that consists of several Gaussian components. It overcomes some of the limitations of Gaussian distribution and can represent more complex data. The GMM distribution can be expressed by the following equation:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \tag{6}$$

where $\boldsymbol{x}$ denotes a potentially multi-dimentional random variable, K is the total number of Gaussian components and π_k is *mixing coefficient* which represents the marginal probability that the data comes from a particular Gaussian component. $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are vectors containing means and variances of a particular Gaussian component.

It can be shown [14] that the log likelihood function for GMM is given by:

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \tag{7}$$

where $\boldsymbol{X}$ is the matrix which row represents a single observation. Due to presence of the sum expression over *k* inside the $\ln$ function, it is impossible to find a simple and elegant analytical solution that would allow us to maximize the log likelihood function. So the cost of using a more complex and sophisticated distribution that allows us to capture multimodal data is the problem that arises when we try to maximize the likelihood function. To solve this problem, an iterative method called Expectation Maximization (EM) algorithm is used.

B. The EM Algorithm

The EM Algorithm is an iterative method which aims in maximising the likelihood function with presence of latent variables, i.e. variables that can not be directly observed. EM can be successfully used for discrete as well as for continuous latent variables $\boldsymbol{Z}$. If we knew directly the complete dataset $\{\boldsymbol{X}, \boldsymbol{Z}\}$, we could easily calculate the model parameters for which the likelihood is maximal. However, the variable $\boldsymbol{Z}$ is latent, thus not observed. The EM algorithm overcomes this limitation and estimates the maximum likelihood, dividing the calculation into two stages:

- E step (Expectation) In this step the model parameters are fixed and we estimate the distribution of latent variables conditioned on fixed observations and parameters.
- M step (Maximisation) This step maximizes the likelihood, assuming the probability distribution of $\boldsymbol{Z}$ calculated in the E step.

The two steps are repeated until convergence.

It should be noted that the EM algorithm is a general solution that can be applied to various models, not only GMMs. In this article we focus on the particular usage of EM algorithm in context of GMMs. In this case the latent variable, often named $\boldsymbol{Z}$, is discrete and can be interpreted as the index of a Gaussian component which generated a particular observation.

The E and M steps of the algorithm for GMMs can be defined and interpreted as following:

- E step
  In this step we try to find the weights, often called *responsibilities*, which can be interpreted as probabilities that the current observation has been generated by the appropriate Gaussian model. From the bayesian perspective, we can view the responsibilities as posterior probabilities that are adjusting the prior probability π_k after observing data $\boldsymbol{x_n}$
  The responsibilities form a $N \times K$ matrix. Each element

$\gamma(z_{nk})$ of the matrix informs us about the probability that the observation n has been generated by the Gaussian model k. With help of Bayes' theorem we can formulate the equation that update the responsibilities using fixed $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ and π_k:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})} \tag{8}$$

It should be noted that the elements of the responsibility matrix form valid probabilities. They are normalized so as that each row sums to 1.

- M step
  The maximisation is done using the responsibilities matrix obtained in the previous step. We forget about the previous model parameters ($\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ and π_k) and override them with newly estimated values. The new values are calculated as following:

$$\boldsymbol{\mu_k} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x_n} \tag{9}$$

$$\boldsymbol{\Sigma_k} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x_n} - \boldsymbol{\mu_k})(\boldsymbol{x_n} - \boldsymbol{\mu_k})^{\mathsf{T}} \tag{10}$$

$$\pi_k = \frac{N_k}{N} \tag{11}$$

where

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{12}$$

After completing both E and M step, the log likelihood should be recalculated using (7). If either the log likelihood or the parameters converged, the algorithm finishes. Otherwise, steps E and M are repeated.

Figure 2 shows the graphical interpretation of the above equations. First (a) the initial means and variances are chosen. Then (b) the responsibilities are calculated – the intensity of blue and red colour informs about the probability that the observation comes respectively from blue and red model. Image (c) shows the maximisation step – the means and variances are recalculated to reflect the current responsibilities. Results obtained after large number of iterations are presented on images (e) and (f).

It can be shown that the EM algorithm guarantees finding the local maximum of the log likelihood[14]. It is an interesting approach that maximises likelihood function and can be successfully applied to models with latent variables, such as GMMs.

C. GMMs' traditional role in speech recognition systems

GMMs are commonly used in ASR systems for modelling emission probabilities in HMM (the probability of an observation given an internal state). In this case GMMs are representing the probability of audio features given a phoneme.

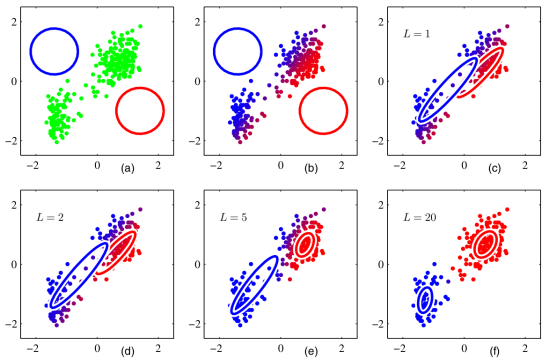

Fig. 2. EM algorithm[14]

The model can be learned using Baum-Welch algorithm - a special case of the EM algorithm. The usage of GMMs in this context has become a standard in speech recognition.

In this article we focus on a different use case for Gaussian mixtures which is feature extraction. The next section gives a justification for GMMs usage in this context and presents an overview of created ASR system.

V. GMM FEATURE EXTRACTION FOR AN ASR SYSTEM

In the speech spectrum we can distinguish several formants (dominant frequencies). The presence of more than one important peaks makes the probability distribution obtained from the speech spectrum multimodal. We need an appropriate mathematical model to handle such a multimodal distribution. Gaussian Mixture Models (GMMs) are a reasonable choice in this situation since they may represent advanced multimodal distributions.

The GMM feature extraction looks as follows:

- Data pre-processing
  Here the methods described in III apply, e.g. data framing, windowing, sampling and DFT. The output of this step is a pre-processed signal in frequency domain.
- Spectral histogram generation
  We convert the audio spectrum to the histogram. With help of the histogram we are able to fit the Gaussian models.
- GMM parameters estimation
  Using the EM algorithm we search for the parameters that maximise the log likelihood of observing data generated according to the spectral histogram. Here the equations described in IV apply with an additional note that a single observation $\boldsymbol{x_n}$ is now one-dimentional since the only variable here is the audio signal in frequency domain. So the GMMs used for formant analysis are univariate ($\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$ are also one-dimentional).
- GMM parameters post-processing
  The parameters obtained in the previous step need to be further processed in aim to achieve a better representation of the speech signal. Here techniques like normalization can be used[7].

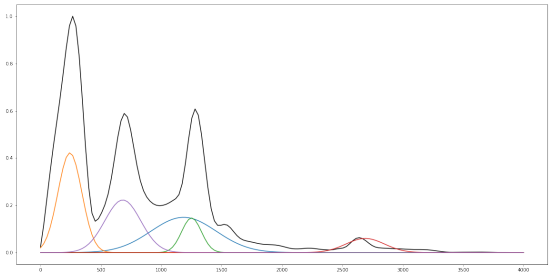

Fig. 3. Speech spectrum with fitted GMMs

The post-processed parameters $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ and π_k can be used as the audio features and treated as input to the classification model (e.g. HMM). Some researches suggest that features obtained using formant analysis are complementary to the MFCC ones[20]. Therefore using a combination of GMM and MFCC features can improve the system's performance.

VI. Results and discussion

Excerpts of *TIDIGITS* database [21] was used for training and evaluation of the presented system. Performance of different feature extraction techniques was measured – GMM and MFCC. For MFCC feature extraction *python_speech_features* library [22] was used. Both results were obtained using the same HMM based classifier. Additionally the proposed work is compared with the isolated words recognition system created in *Matlab* using MFCC feature extraction, presented in [23].

Table I presents achieved accuracy for each of the discussed systems.

TABLE I
Comparison of the systems accuracy

System name	Accuracy (%)
GMM based system	36,6
MFCC based system	98,7
Matlab system	43.08

The score achieved by the GMM based system is significantly lower than the MFCC based system, in which 13 of the MFCC features are used. There are several factors that can explain the difference between these scores. Firstly, 10 GMM features were being extracted for each frame, whereas the MFCC method was generating 13 of them. Further, the implemented EM algorithm was simplified, which caused the GMM features to be correlated and thus containing less information. The vulnerability of the first step of the EM algorithm could lead to a situation, where the same initial means were chosen and thus the features were duplicated. Another aspect of the proposed system that could be potentially improved is the GMM features post-processing. The discussed results were affected by the lack of feature normalization.

The above improvements could potentially increase the system's accuracy. Additionally, a combination of GMM and MFCC features could achieve the best performance, as suggested in[20].

References

[1] V. Gupta, J. Bryan, and J. Gowdy, "A speaker-independent speech-recognition system based on linear prediction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 27–33, 1978.

[2] S. A. Alim and N. K. A. Rashid, "Some commonly used speech feature extraction algorithms," *intechopen*, 2018.

[3] S. Suksri, "Speech recognition using mfcc," 09 2015.

[4] G. Vyas and B. Kumari, "Speaker recognition system based on mfcc and dct," *International Journal of engineering and advanced technology*, vol. 2, pp. 167–169, 06 2013.

[5] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *J Comput*, vol. 2, 03 2010.

[6] P. Zolfaghari and T. Robinson, "Formant analysis using mixtures of gaussians," 11 1996, pp. 1229 – 1232 vol.2.

[7] M. N. Stuttle, "A gaussian mixture model spectral representation for speech recognition," 07 2003.

[8] Z. Zhang, "Mechanics of human voice production and control," *The Journal of the Acoustical Society of America*, vol. 140, pp. 2614–2635, 10 2016.

[9] "Phones and phonemes," *P. Coxhead*, 2006.

[10] A. Bizzocchi, "How many phonemes does the english language have?" *International Journal on Studies in English Language and Literature (IJSELL)*, vol. 5, pp. 36–46, 10 2017.

[11] D. J. Broad, "Formants in automatic speech recognition," *International Journal of Man-Machine Studies*, vol. 4, no. 4, pp. 411 – 424, 1972. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020737372800373

[12] N. Najkar, F. Razzazi, and H. Sameti, "A novel approach to hmm-based speech recognition systems using particle swarm optimization," *Mathematical and Computer Modelling*, vol. 52, no. 11, pp. 1910 – 1920, 2010, the BIC-TA 2009 Special Issue. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895717710001597

[13] J. S. Bridle, "Neural networks or hidden markov models for automatic speech recognition: Is there a choice?" in *Speech Recognition and Understanding*, P. Laface and R. De Mori, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 225–236.

[14] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[15] R. Hokking, K. Woraratpanya, and Y. Kuroki, "Speech recognition of different sampling rates using fractal code descriptor," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp. 1–5.

[16] B. B. Monson, E. J. Hunter, A. J. Lotto, and B. H. Story, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, p. 587, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00587

[17] C. Sunitha and E. Chandra, "Speaker recognition using mfcc and improved weighted vector quantization algorithm," *International Journal of Engineering and Technology*, vol. 7, pp. 1685–1692, 11 2015.

[18] C.-P. Chen and J. Bilmes, "Mva processing of speech features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, pp. 257 – 270, 02 2007.

[19] K. S. Rao and K. E. Manjunath, *Speech Recognition Using Articulatory and Excitation Source Features*, 1st ed. Springer Publishing Company, Incorporated, 2017.

[20] J. Holmes, W. Holmes, and P. Garner, "Using formant frequencies in speech recognition." 01 1997.

[21] *TIDIGITS - Linguistic Data Consortium - LDC Catalog*, 1993 (accessed December, 2020). [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S10

[22] accessed January, 2021. [Online]. Available: https://github.com/jameslyons/python_speech_features

[23] H.-H. Le, *GMM-HMM (multiple Gaussian) for isolated words recognition*, 2017 (accessed December, 2020). [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/64297-gmm-hmm-multiple-gaussian-for-isolated-words-recognition