CONTENT BASED AUDIO EXTRACTION: A COMPARATIVE STUDY OF VARIOUS FEATURE EXTRACTION TECHNIQUES

Prof. Mandar S. Joshi, Miss. Rameeza I. Lambey, Miss. Urvi N. Pokar, Mr. Nainish N. Kher
Asst. Professor, Student, Student, Student
Information Technology,
Finolex Academy of Management and Technology, Ratnagiri, India

Abstract : This paper introduces the various feature extraction techniques from audio. It first gives the idea about the different types of noise that can be encountered in an audio. Then the two end-point detection algorithms are discussed. Then it specifies different feature extraction algorithms and their comparison. Finally, it gives the brief description about the pattern classification algorithms commonly used in speech recognition systems.

***IndexTerms* - Feature Extraction, Zero Crossing, Mel-Frequency Cepstral Coefficient, Linear Predictive Coding, Support Vector, Hidden Markov Model.**

I. INTRODUCTION

In recent years, with the rapid development of the computer, the computing efficiency is improved. However, the traditional operation of human-computer interaction has been unable to meet the needs of users, so developers began to seek more convenient human-computer interaction mode. With the rapid development and popularity of smart phones, users want to replace the keyboard by voice to more convenient for human-computer interaction, to achieve a direct dialogue between human and machine. In order to meet the needs of users, many researches are going on. Due to the convenience provided by such voice commands over keyboard or touch commands, audio recognition technology has become a very competitive technology. [11]

Here, the term audio refers to generic sound signals, which include speech, dialogue, music, songs, radio broadcasts, audio-tracks of video programs, noise and mixture of any of these. Common features of audio include – time-domain features (Zero Crossing Rate, short time energy), frequency-domain (spectral) features (Pitch, sub-band energy ratio), psycho-acoustic features (Four-Hz modulation energy, spectral roll-off point). [4]

Audio feature extraction serves as the basis for a wide range of applications in the areas of speech processing, multimedia data management and distribution, security, biometrics and bioacoustics. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. [6]

Some of the audio features that have been successfully used for classification of audio signals include MFCC (Mel-Frequency Cepstral Co-efficient), Spectral Similarity, Timbral Texture, MPEG-7 descriptors, Zero Crossing Rate, entropy, compactness, spectral flux, octaves and band periodicity. Among them, MFCC is inarguably the single most important feature that is widely used for the classification purpose. [3]

II. PRE-PROCESSING OF SPEECH

A) Analysis of noise contents:

Additive noise is additive to the speech signal in both, the power spectrum domain and the autocorrelation domain. Channel disturbance on the other hand is multiplicative in nature in the autocorrelation domain. Thus, the technique to remove additive noise is to subtract in either the power spectrum domain or the autocorrelation domain. The technique to remove channel distortion however, involves a more complicated set of steps. [5]

The multiplicative nature of the distortion can be converted into an additive nature by shifting the speech from the autocorrelation domain to the logarithmic domain. Here, a filtering technique is applied in which mean subtraction is used to remove the channel effect.

The noises can be broadly classified as channel distortion and additive noise. The additive noise is coloured in nature. The coloured noises that have been considered are babble noise, factory noise and F-16 noise. The channel distortion is in the form of a random sequence of numbers emulating a Gaussian Channel. [5]

B) End-Point Detection and Segmentation:

The speech signal *x* is taken and a pre-emphasis filter is used to perform the task of enhancing the dominant parts of the signal. The emphasized speech signal is the divided into frames and the framed signal is then used as an input to the end point detection algorithm.

Following are the two popular algorithms used for extracting the content by detecting the end-points of the spoken data.

i) ZCR-EN (Zero Crossing Rate - short time Energy):

This algorithm uses two time domain parameters to decide the boundary between silence voice components. ZCR is zero crossing rate which is defined as the number of times that a signal changes signs in a particular frame and can be calculated using (1).

$$\frac{1}{2 \cdot N} \cdot \sum_{m} | \operatorname{sgn}(x(m)) - \operatorname{sgn}(x(m-1)) |$$

$$1 \leq n \leq N : 1 \leq m \leq M \qquad (1)$$

where *M* is the number of samples per frame and *N* is the total number of frames in the signal. Short time energy is defined as the sum of the squares of the magnitudes of the samples taken per frame. It can be calculated using (2).

$$\mathrm{E_n} = \frac{1}{N} \cdot \sum_{m} x(m)^2$$

$$1 \leq n \leq N : 1 \leq m \leq M \qquad (2)$$

The algorithm to find the endpoints using these to parameters has been presented below:

First, threshold values on the basis of which a certain frame is accepted or rejected are calculated. The zero crossing rate threshold OCRZ is determined using (3).

$$\mathrm{OCRZ} = \frac{1}{N} \Sigma_{i=1}^{N} x_i \times 2\sqrt{\frac{1}{n-1} \Sigma_{i=1}^{N} (x_i - \overline{x})^2} \qquad (3)$$

The upper (TUL) and lower (TLL) thresholds of energy are calculated using (4).

TLL = min(0.03 × (*IMX* – *IMN*) + *IMN*, 4 × *IMN*)
TUL = 5 × *ITL* (4)

where, IMN and IMX are the minimum and maximum energy levels found in the signal.

Searching is started from the beginning of the framed signal until the energy crosses TUL. Then the search is backed off towards the signal beginning until the first point at which the energy falls below TLL is reached. This is marked as the provisional beginning point - N1. N2 (the end point) is evaluated in a similar way. Again, the signal is searched from the beginning and ZCR is examined. If this measure exceeds the OCRZ threshold 3 or more times, N1 is moved to the first point at which the threshold is exceeded. N1 is defined as the formal beginning point. Formal endpoint N2 using OCRZ is evaluated in a similar manner. [5]

ii) VFR (Variable Frame Rate):

Variable frame rate (VFR) is a technique used for discarding frames that are too much alike. The method emphasizes the transient regions, which are more relevant for speech recognition. The algorithm consists of three steps. At first, the speech signal corresponding to a single word is pre-processed and the background noise is estimated which is used to decide the threshold values for the following steps. In the second step, the starting-point and the ending-point of the voiced sound are located to be used as reference endpoints based on time domain features of short time energy and zero-crossing rate. And finally, the accurate endpoints of the utterance are located according to the frequency parameter called mel-frequency cepstrum of the sequence of speech signals between the reference endpoints. [5]

$$E_b = \begin{cases} \dfrac{E_{k-1} + E_K}{2} & if\ 0.5 \leq \dfrac{E_{k-1}}{E_k} \leq 2 \\ \min(E_{k-1}, E_k); & \text{otherwise} \end{cases} \qquad (5)$$

For determining background noise ZCR, (6) is used.

$$Z_N = \begin{cases} \dfrac{Z_f + Z_b}{2} & \text{if } 0.5 \leq \dfrac{Z_f}{Z_b} \leq 2 \\ \text{rejected}; & \text{otherwise} \end{cases} \qquad (6)$$

$$Z_f = \begin{cases} \frac{Z_1 + Z_2}{2}; & \text{if } 0.5 \le \frac{Z_1}{Z_2} \le 2 \\ \min(Z_1, Z_2); & \text{otherwise} \end{cases}$$

$$Z_b = \begin{cases} \frac{Z_{k-1} + Z_k}{2}; & \text{if } 0.5 \le \frac{Z_{k-1}}{Z_k} \le 2 \\ \min(Z_{k-1}, Z_k); & \text{otherwise} \end{cases} \qquad (7)$$

Energy threshold T_E and ZCR threshold T_Z are calculated using the background energy and ZCR levels of E_N and Z_N. The energy function is searched and the first frame whose energy is above *TE***,** is assumed to be the starting point as in (8).
$P_{F2} = k\ \{E_k > T_E;\ k = 1, 2, ..., K\}$ (8)
where E_k is defined by (2).

The energy function is then searched backwards from right to left, the ending-point of the voiced sound is obtained by:
$P_{B2} = k\ \{E_k > T_E;\ k = 1, 2, ..., K\}$ (9)

The zero-crossing parameter is then used to relax the endpoints. The zero-crossing function is searched from point P_{F3} backwards to obtain
$P_{F2} = k\ \{Z_k > T_Z;\ k = P_{F2}, P_{F2-1}, ..., 1\}$ (10)
where Z_K is defined in (1).

The zero-crossing function is searched from point P_{B3} forwards, to obtain
$P_{B2} = k\ \{Z_k > T_Z;\ k = P_{B2}, P_{B2+1}, ..., K\}$ (11)

$D(i, j)$, the Euclidean distance between the current frame i, and the last retained frame j is evaluated using mel-frequency cepstrum coefficients of the signal. Euclidean threshold *TD* is experimentally derived to be -5.8. The Euclidean function is searched forward from P_{F2}
$P_{F1} = k\ \{D(k, k + 1) > T_D\ \&\&\ D(k, k + 2) > T_D\ \&\&\ D\ (k, k + 3) > T_D\}$ (12)

The Euclidean function is the searched from P_{B2} backwards
$P_{B1} = k\ \{D(k, k - 1) > T_D\ \&\&\ D\ (k, k - 2) > T_D\ \&\&\ D(k, k - 3) > T_D\}$ (13)
The points finally obtained are the actual endpoints.

III. Feature Extraction Techniques

In automatic speech recognition system, the important step is to extract features which is used to identify the important contents of speech signal. It is also used to identify the unwanted noise present in nearby environment. Common features of audio are as follows –

- **Magnitude spectrum:** This feature is calculated by initially calculating the FFT with the Hamming window. The value is calculated by adding the squares of real and the imaginary parts of the audio parameters of each bin.
- **FFT Bin Frequencies**: It is calculated by using bin level (in Hz), which is used in power or magnitude spectrum and then the calculations is done by FFT of the window size which is provided to the feature extractor.
- **Power spectrum**: This feature provides the distribution of audio signals amongst different frequencies.
- **Compactness:** This is calculated by finding sum of all values in Bit Histogram.
- **Bit Histogram**: It shows the strength of different measured periods of the signal. It is calculated by Root Mean Square of 256 window and taking FFT of result.
- **ZCR**: This feature is calculated by finding the number of the times signal changes from one sample to another (reaches the Zero axis) [2].

Audio feature extraction also plays important role in examining and predicting audio components. Different applications required for efficient feature extraction techniques are: Auditory scene analysis, content-based audio retrieval, indexing of audio. Some of the audio feature extraction techniques include: Most popular are LPC and MFCC. [9]

i) LPC (Linear Predictive Coding):

It is a techniques used for speech recognition to evaluate basic speech stipulations like pitch, formant and spectral envelope of speech signal in compressed form. It is a technique used for encrypting better quality speech at low bit rates. It calculates the power spectrum of the signal. This techniques provides basic speed parameters, which can be used for efficient computations of performance evaluation. The variation depends on pitch, formant, intensity, frequency. [9]

ii) MFCC (Mel-Frequency Cepstral Coefficient):

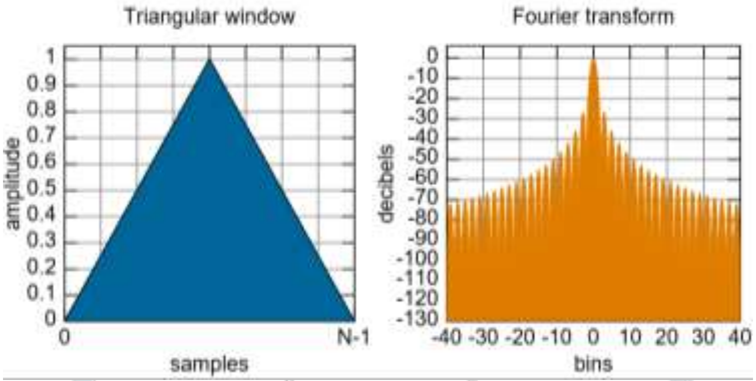
It is used in audio and speech processing. They are derived from a type of cepstral (result of taking the IFT of logarithm of spectrum of signal) representation of audio clip. The difference between the mel-frequency cepstrum and cepstrum is the frequency bands are equally spaced on the mel-scale (listeners to be in distance from each other), which approximates the human auditory range more closely than the linear-spaced frequency bands. [9]

MFCCs are commonly used in feature extraction techniques such as systems which can automatically retrieve numbers from voice call as well as audio similarity measures.

Derivation:

a) Take the FT

b) Map the powers of the spectrum obtained above onto the mel-scale, using triangular overlapping windows (2nd order of B-spline window).



c) Take the logs of powers of each of the mel-frequeries.

d) Take the discrete cosine transform of the mel log powers.

e) The MFCCs are the amplitudes of the resulting spectrum.

Advantage of MFCC:

As it is less complex in implementing the feature extraction algorithm, only sixteenth coefficient of MFCC related to mel-scale frequency of speech spectrum are extracted from spoken word samples in database. [9]

iii) LPCC (Linear Predictive Cepstral Coefficient):

In LPCC before converting into cepstral coefficient, coefficient of LPC spectral envelope id made. The steps of LPCC are similar to MFCC techniques. In LPCC smoothing of spectrum is done by auto aggressive filter. In the next step inverse Fourier transform is applied to the power spectral density (PSD) to obtain the autocorrelation function. After the analysis of LPC, input signal passes through the filter and provides output as a power spectrum [11].

Segmentation:

In many applications, audio segmentation is the process of partitioning a continuous audio stream in terms of phonic, homogeneous or equal category regions. The goal is to find the phonic changes in an audio signal [2]. Therefore, useful information is produced such as dividing into speaker signals and caller identities, allowing for the automatic indexing and retrieve the information of all the occurrences.

Different approaches used for segmentation are:

a. Energy Based segmentation.
b. Model Based segmentation
c. Metric Based segmentation.

IV. Pattern Classification Algorithms

i) Minimum Distance Classifier (MDC):

In speech recognition or STT conversion there are mainly two phases first is training phase and second one is testing phase. For classification, during training phase zero crossing points (ZCP) corresponding to the different words are pre-computed and stored as reference ZCPs [10]. Minimum distance classifier computes Euclidean distance between the zero crossing points of the uttered word and zero crossing points of words from database. The word having least Euclidean distance is declared as uttered word.

Euclidean distance is given as:

$$d^2(\bar{x}, \bar{p}) = \sum_{1}^{k} (x_i - p_i)^2$$

where, *x* and *p* ZCP database. i.e. *x* is a ZCP vector of uttered word. *p* is a ZCP vector of different words. i varies from 1 to k (i.e. no. of ZCPs of a particular word). The sum of squares of the difference between the individual zero crossing points is computed to calculate the Euclidean distance i.e. distance between the uttered word and all words in the database is found out. The word in the database with least distance is declared as the uttered word. [10]

ii) Support Vector Machine (SVM):

SVM is one of the effective method of pattern classification. SVM use linear and nonlinear separating hyper-planes for data classification. First input is mapped into a high dimensional space and then with the help of hyper-plane it distinguishes the classes.

The inner product, kernel which is caused by the high dimensional mapping is a crucial aspect of opting SVMs successfully i.e. a high dimensional feature space is implicitly introduced by a computationally efficient kernel mapping and in a high dimensional feature space SVM finds a separating surface with a large margin between training samples of two classes . And large margin implies a better generalization ability. SVM uses discriminative approach. The classification of any fixed length data vectors is possible by SVM. It cannot be readily applied to task involving variable length data classification. [10]

The support vector classifier uses the function:

$$f(x) = ([\alpha * K_s(x)]) + b$$

Where, $K_s(x) = [k(x, s_1), \ldots\ldots k(x, s_d)]^T$ is a vector of evaluation of kernel functions centered at the support vectors.

$$f(x) = ([\alpha * K_s(x)]) + b$$

Which are usually subset of the training data.

The classification rule is defined as:

$$q(x) = \{1 \, forf(x) \geq 0$$

$$\{2 \, forf(x) < 0$$

And multiclass classification function and rule is defined as:

$$f_y(x) = (\alpha_y * k_s(x)) \;+ b_{y,}\, y \in Y$$

$$Q(x) = \arg\max f_y(x), y \in Y$$

iii) Dynamic Time Wrapping (DTW):

Any two time series can be varied in time and speed which is called wrapping points. Thus, data time wrapping technique is one of the most used feature matching (i.e. classification) techniques. Consequently, this technique is used to find the optimal alignment between two time series, as well as, measuring the similarity between those time series (Zhang etal., 2013).

The DTW is employing linear time wrapping by comparing signals of two time series based on linear mapping of the two temporal dimensions (Chapaneri, 2012). Thus, DTW allow non-linear alignment of one signal to another by minimizing the distance between two signals. Therefore, this wrapping can be used to extract figure recognition based on the similarity and dissimilarity between those signals. From speech signals point of view, the duration of each spoken word or digit can vary but the overall speech waveform are similar for same word or digit. Therefore, by applying the DTW technique the corresponding regions between the two time series can be extracted easily to be used in matching processes (Muda et al., 2010).

In more details, Chapaneri, S. measured the optimal wrap path for a given two time series A and B where the length of A is |A| and the length of B is |B|. Thus, A = A1, A2, A3… A|A|, and B = A = B1, B2, B3… B|B|, a wrap can be constructed W = W1, W2, W3 … W k. Where k is the length of wrap path for a kth element that is max (|A|, |B|) <= k <= |A| + |B|, and W k = (i, j) where i is the length of time series A, and j is the length of time series B. Consequently, the optimal wrap path that represents the minimum distance between two time series can be calculated using (equation 14) (Chapaneri, 2012). [1]

$$Dist(w) = \sum_{k=1}^{|k|} \frac{Dist(w_{ki}, \mathrm{w}_{kj})}{(|A| + |\mathrm{B}|)} \qquad (14)$$

iv) Hidden Markov Model (HMM) technique:

The core idea in using HMM for speech recognition applications is to create a stochastic models from known utterances and compares it with the unknown utterances was generated by speaker. An HMM M is defined by a set of states N that have K observation symbols as well as, three possibility metrics for each state which are in (equation 15) (Ghahramani, 2001)..

$$M = \{\Pi, A, B\} \qquad (15)$$

where:

- Π: initial state probability.
- A: at,j state transition probability.
- B: bt, j, k symbol emission probabilities.

For each HMM system, it could be use three different types of topologies to employ Markov chain which are ergodic model, general left to right model, and linear model. Figure 1 illustrates HMM topologies for a system with four states. Consequently, each state has its own probability which leads to compute the probability for an occurrence of state in a given situation of another state using Bayesian rule.

In this context, for any system employs HMM technique three basic algorithms which are classification, training, and evaluation algorithms. In classification algorithm, the recognition process is enabled for any unknown utterance by identifying the unknown observations sequence via choosing the most likely class to have produced the observation sequence. In training algorithm, the model is responsible to store data collected for a specific language (i.e. in our research the language was the English language). In the evaluation algorithm, the probability of an observation sequence is computed for matching processes.

Fig. 1. HMM three states topologies for a system with four states 1, 2, 3, and 4 (Paul, 1990)

The classification algorithm was employed for a given observations O = O1, O2, O3 … OT. A chosen class was computed using (equation 16) (Paul, 1990).

$$\text{Chosen\_Class} = \arg MAX\left[P\left(M_{class} \mid O\right)\right] \qquad (16)$$

Therefore, by applying Bayesian rule to find the probability was computed using (equation 17) (Paul, 1990).[5]

$$P\left(M_{class} \mid O\right) = \frac{P(\mathrm{O} \mid M_{class})\,\mathrm{P}(\mathrm{M}_{class})}{P(O)} \qquad (17)$$

V. RESULTS AND DISCUSSION

Table1: Comparison of Various Speech Recognition Techniques [7]

Technique	Description	Advantages	Disadvantages	Applications
DTW (Dynamic Time Wrapping)	• It is a statistical approach which is used to recognize the speech. • Its main principle is to compare two dynamic patterns & measure its similarity by calculating a minimum distance between them.	• It is powerful for measuring similarity between two time series which may vary in time or speed. • The training procedure is very simple & fast.	• The main problem is to prepare reference template. • Single template is not sufficient.	It is used in small scale embedded speech recognition system those embedded in cellphones.
SVM (Support Vector Machine)	• It is simple & effective method for classification of speech recognition. • It is a binary non-linear classifier capable of guessing whether an input vector x belongs to a class1 or class2 category.	• Minimize the structural risk. • Increase the robustness of the system. • Training is relatively easy. • Local optimality is not needed. • It scales relatively well for high dimensional data.	• Good kernel function is needed. • Requires full labelling of input data. • It is only applicable for two class tasks.	It is used for speaker & language recognition. They are helpful in text & hypertext categorization.
HMM (Hidden Markov Model)	• It is a mathematical framework or statistical model of a sequence of feature vector observations. • State sequences are hidden & the observations are	• It is fast in its initial training. • Performs quite well in noisy environment.	• Large prior modelling, assumptions about data have to make. • Amount of data & no. of parameters that need to be set during training is	It has been used in low level NLP processes such as phrase chunking, extracting necessary information from documents & part of speech tagging.

	probabilistic functions of the state.		used. • It doesn't minimize the probability of observation of instances from other classes.	

Comparison between MFCC and LPC:

- In MFCC, filter bank analysis is used for computations while in LPCC Linear predicative coding is used for computation.
- MFCC is better than LPC.
- LPC focuses on formant structure.
- MFCC is non-linear while LPC is linear.
- MFCC is computationally complex compared to LPC as MFCC contains many steps to compute the coefficients and LPC contains few steps for computations.
- System efficiency:
  MFCC and HMM: 90.67%
  LPC and HMM: 80.67%

VI. Acknowledgment

We thank our college Finolex Academy of Management and Technology for allowing us to publish this technical research paper. We thank our project guide Prof. Mandar S. Joshi sir who provided insight and expertise that greatly assisted the research. We thank our librarians for helping us with the material required as the reference. Any errors are our own and should not tarnish the reputations of these esteemed persons.

References

[1] Dr. Hebah H. Nasereddin and Ayoub Abdel Rahman Omari, "Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation", Computing Conference 2017, 18-20 July 2017, London, UK.

[2] Gayatri M. Bhandari, "Different Audio Feature Extraction using Segmentation", International Journal of Research in Science and Technology, 09-February 2016, Volume 2.

[3] Guihua Wen, Jian Tuo, Lijun Jiang and Jia Wei, "Audio Feature Extraction for Classification using Relative Transformation", IEEE ICALIP2012.

[4] Hualu Wang, Ajay Divakaran, Anthony Vetro, Shih-Fu Chang and Huifang Sun, "Survey of Compressed-domain Features and in Audio-visual Indexing and Analysis", Journal of Visual Communication and Image Representation, 21st February 2003, Elsevier Science (USA).

[5] Kapil Sharma, H. P. Sinha and R. K. Aggarwal, "Comparative study of Speech Recognition System using various Feature Extraction Techniques", International Journal of Information Technology and Knowledge Management, July-Decemeber 2010, Volume 3, No. 2, pp. 695-698.

[6] Karthikeyan Umapathy, Shridhar Krishnan and Raveendra K. Rao, "Audio Signal Feature Extraction and Classification using Local Discriminant Bases", IEEE Transactions on Audio, Speech and Language Processing, Vol. 15, No. 4, May 2007.

[7] Neerja Arora, "Automatic Speech Recognition System: A Review", International Journal of Computer Applications, Volume 151 - No.1, October2016.

[8] Rajiv Chechi, Reetu," Performance Analysis of MFCC and LPCC Techniques in Automatic Speech Recogntion", International Journal of Engineering Research and Technology, 9 September 2013, Volume 2.

[9] Urmila Shrawankar,Dr.Vilas Thakare, "Techniques for Feature Extraction in Speech Recognition System: A Comparative Study", SGB Amravati University.

[10] Yogita H. Ghadage and Sushama D. Shelke, "Speech to Text Conversion for Multilingual Languages", International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[11] Zhengzheng Liu, Cong Li and Sanxing Cao, "Audio Fingerprint Extraction Based on Time-Frequency Domain", 2016 2nd IEEE International Conference on Computer and Communications.