Optimal Feature Extraction and Selection Techniques for Speech Processing: A Review

Ankita N. Chadha, Mukesh A. Zaveri and Jignesh N. Sarvaiya

***Abstract*—Today speech processing is one of the demanding applications among all others. This article highlights two important aspects of speech processing, namely which feature representation must be employed and what is their selection criteria. Depending on different application areas, speech processing needs different set of features and techniques to extract them. At the same time it is necessary to choose optimal set of features or optimal size of feature vector in order to reduce timing complexity. In this context we address various features suitable for different speech processing and how these features should be selected are discussed at the length. There is a need to reduce the size of feature vector, thus various techniques for selecting optimal set of features are also discussed in this article. This paper helps a reader to get a quick insight in the area of speech processing.**

Keyword: *Feature extraction, LPC, Complex Cepstrum, MFCC, HNM, aHM, Feature Selection, PCA, LDA, CCA*

I. Introduction

Speech and audio signals are produced due to variations in pressure and displacement of air falling on perception device called ears. Such physical signals need to be treated and processed in order to make them easier for storage and machine learning. This leads to analysis of these audible sounds through unique representation methodologies termed as feature extraction[1].

Feature designing are early phase in any speech processing application such as speech recognition[1], speaker recognition[2], speech enhancement[3], speech transformation[4], emotional speech recognition[5] and audio retrieval[6]. All of these speech areas need very precise parameterization in order to build robust systems. The main work of this paper is two fold:

i. Presenting an extensive survey of various features across various fields of speech processing including audio and emotion as well.
ii. Highlighting the most popular feature selection algorithms that are necessary for improving performance of machine learning task.

The paper is organized as follows: section 2 describes various feature extraction techniques and section 3 is dedicated to feature selection methods. Lastly the conclusion of this work.

A.N. Chadha is working as a research scholar in Department of Electronics Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, India 395007, email: ankitaism@gmail.com

M.A. Zaveri is working with Department of Computer Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, India 395007, email:mazaveri@coed.svnit.ac.in

J.N. Sarvaiya is working with Department of Electronics Engineering, Sardar Vallabhbhai National Institute of Technology, Surat, India 395007, email: jns@eced.svnit.ac.in

II. Feature Extraction Techniques

In any data analysis process, feature construction is the key stage before applying any data learning algorithm. Features signify the most distinct traits drawn out of speech signal which comprises of pauses, silences and a lot of other redundant information. A wide range of feature representation techniques have evolved over the period due to rise in market demands and faster processing and transmission needs. This work presents a review of quite a many features and their categorical domains during 1970 to till date.

Speech signals emerged as a highly non-stationary signals which were not only difficult to observe but also more noise prone. The main aim of any feature extraction technique is to preserve the variations and relevant information of speech signal which are compact and computationally inexpensive.

Based on biological categorization, features may be production and perception based; considering domain of processing, there may be temporal, frequency, cepstral and eigen domains[6]. The temporal features have amplitude, power and zero crossing rate, primarily help in taking voicing decisions. The frequency domain features are the oldest and most popular type, comprising of physical features which are statistical based and perceptual features which deal with semantic meaning from listener's perspective. These perceptual features also have brightness, tonality, loudness and pitch as shown in figure. Then, the cepstral features are sub-divided into perceptual filter banks, advanced auditory model and auto-regression features.

A. Linear Prediction Coefficients (LPC)

The linearly coded parameters are well known for their accurate parameter estimation and simplest feature representation[7]. The current speech sample is predicted from past speech samples and excitation signal from lungs as shown in Figure 1. It is basically an all pole system with gain G.

$$s_n = \sum_{l=1}^{P} a_l s_{n-l} + G e_n \tag{1}$$

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{l=1}^{P} a_l z^{-l}} \tag{2}$$

The LPC based features suffer from issues of quantization, stability and interpolation. Further more, it assumes the glottal excitation to be independent of vocal tract filter. The issue of quantization can be overcome using Line spectral frequencies (LSF) based features[8]. The LSF features quantize all pole

978-1-5090-0396-9/16/$31.00 ©2016 IEEE

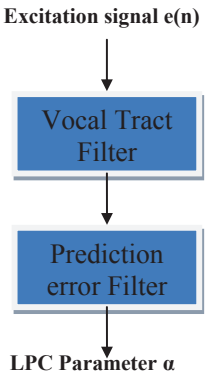

Fig. 1. LPC model

filter polynomial into symmetrical and anti symmetrical equations and its reverse is also possible. The LPC based features have their equivalents called Log Area Ratio and Reflection coefficients. The LP based features are employed in almost all the speech processing applications filtering, recognizers and synthesis tasks.

B. Complex Cepstral Features (CC)

The cepstral based analysis gained popularity due to its elegance in smoothly separating glottal excitation and vocal tract parameters through process of liftering from speech signal [4]. The cepstrum is obtained by taking Inverse DFT of logarithm of magnitude of DFT of a speech sample. Since domain is no longer in time or frequency it is termed as quefrency domain. When we take magnitude, the cepstrum is real and when magnitude along with phase is considered, the cepstrum is complex as shown in Figure 2.

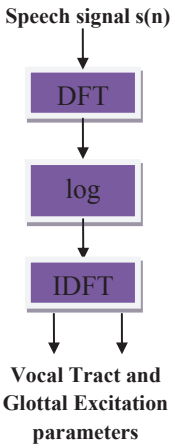

Fig. 2. Complex Cepstrum Model

$$c_l = IDFT\left[log\left[DFT(s_n)\right]\right] = IDFT\left[log\left[DFT(g_n * v_n)\right]\right] \quad (3)$$

where, s_n speech frame at n^{th} time interval The cepstral features have a comparatively higher filter order in contrast to LP based features leading to computational complexity. These are widely used in speech enhancement and voice transformation[3], [4].

C. Mel-frequency cepstral coefficients (MFCC)

The mel frequency is susceptible to human auditory system. Thus a filter bank which linear upto 1000 Hz and nonlinear beyond that is called Mel filter bank. The MFCC features employ Discrete cosine Transform after logarithm operation in real cepstrum as shown in Figure 3. Such a feature method yields very high accuracy in recognition and classification tasks; but discards pitch and phase related information[9]. Hence is it not very desirable for transformation and synthesis tasks[10].

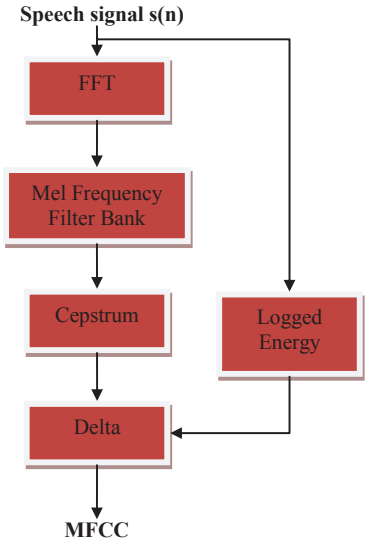

Fig. 3. MFCC Model

D. Harmonic Noise Models (HNM)

These models are extensions to sine model by Stalianou[11]. These include noise component for unvoiced sounds and are widely used in pitch synchronous based speech concatenation approach. These models work well at lower sampling rates and where speech signal is stationary which is usually not the case.

E. Adaptive Harmonic Model (a-HM)

These models are adaptive in the sense that they employ an adaptive iterative refinement algorithm for calculating sinusoidal components. This model when compared with other feature extraction outperforms in terms of perception test as the speech is highly intelligible[12].

F. Jitter and Shimmer

Jitter may be defined in terms of absolute, relative and rap values by measuring variations in fundamental frequency, f_0 from one cycle to another[13].

$$Jitter = \frac{1}{N-1}|t_i - t_{i+1}| \tag{4}$$

where N is the total number of f_0 periods extracted, t_i are i^{th} time instant at f_0 period lengths.

Shimmer is a measure of amplitude variations from peak-to-peak at instants of f_0[13].

$$Shimmer(dB) = \frac{1}{N-1}\sum |20log(A_{i+1}/A_i)| \tag{5}$$

where A_i is the peak to peak amplitude at f_0 period.

These features are amongst features used to represent emotions in speech and speech recognizers as well. These are difficult to extract as the extraction solely depends of how accurately the fundamental frequency has been obtained.

In order to have an overview of features for various applications of speech processing, the below Table I is shown.

TABLE I
SPEECH PROCESSING APPLICATIONS AND THEIR RESPECTIVE FEATURES

	Applications	**Features**
1	Speech enhancement	Cepstrum, MMSE, Spectral subtraction[3], sub-space algorithm, Weiner and Kalman Filtering
2	Speech Recognition	LPC, Cepstrum, MFCC, PLP[14], f0, LPCC, FFT, RASTA[15], i-vectors, PLDA
3	Speaker Recognition	LP residual, LPC, harmonic features, formant frequencies, pitch contours, co-articulation, Zero crossing rate, energy, MFCC, PLP
4	Audio information Retrieval	Pitch, timbral features, zero crossing, centroid, roll-off, MFCC, rhythm features, MPEG-7, time envelope, root mean square, low energy rate, loudness[6].
5	Emotion Recognition	LPC, MFCC, PLP, LPCC, pitch, Teager Energy Operator[16], LFPC, energy, mean, variance, pitch contour, duration, intensity, Jitter, Shimmer
6	Voice Transformation	LPC, LSF, Complex Cepstrum, Wavelet features, Salient sub-band[17], Filter-banks[18], HNM, a-HM, mcep-f0, LP-residual.

III. Feature Selection Techniques

In real-world classification/mapping problems, size of a dataset is very large . Thus the learning might not work as well before removing these unwanted features. One can reduce the number of features and hence avoiding over-fitting issues. With the creation of huge databases and the consequent requirements for good machine learning techniques, new problems arise and novel approaches to feature selection are in demand [19].

The prime task of any feature extraction algorithm is general data reduction for improving performance of system and data understanding. Assessing the quality of selected features is a critical task in any feature selection methodology. Feature selection techniques are categorized as supervised: determines feature relevance by evaluating feature's correlation with the class, and without labels and unsupervised techniques: exploits data variance and separability to evaluate feature relevance. Another categorization of feature selection may be flat, streaming and structured as depicted in Figure 4. The flat feature selection are further sub-divided into filter, wrapper and embedded models while structured models are classified as graph, tree and group selection methods [20].

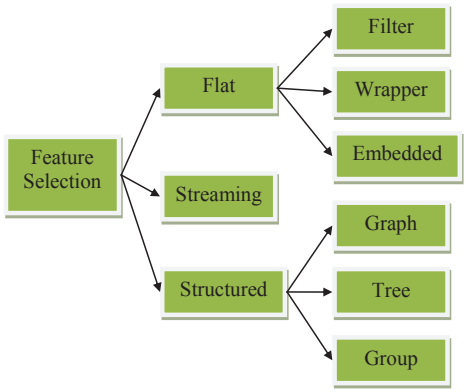

Fig. 4. Feature Selection categorization

A. Analysis of Variance One versus All (ANOVA)

One of the most efficient and time saving feature selection method is ANOVA [21]. The probability values (p-values) are used for each feature to rank them in ascending order. However, it goes by the assumption of equal variances and normality amongst all features along with independent evaluation of features. These concerns sometimes lead to smaller p-values being allotted to highly correlated features and accordingly are selected in their subsets. Yet, ANOVA is widely used as a pre-selection algorithm for various other dimensionality reduction techniques such as LDA (Linear Discriminant Analysis) and SVM-RFE (Support Vector Machine). This pre-selection works well only if selected size of feature vectors are large enough in order to avoid any loss of important features.

B. Principal Component Analysis (PCA)

It is an unsupervised feature selection algorithm that projects variables orthogonally into a new space, according to their variances. The features with lower variances are ignored [22]. The PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the important information from the data able and to express this information as a set of new orthogonal variables called principal components. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in maps.

$$X = P\Delta Q^T Q \tag{6}$$

where X is SVD speech matrix, P= PCA components and Q is linear combination matrix to compute factor scores F [22].

C. Linear Discriminant Analysis (LDA)

The LDA technique is used to reduce dimensionality while preserving as much of the class discriminatory information as possible[23].

$$y = w^T x \tag{7}$$

where y is a scalar projecting the samples x onto a line with w as the class matrix.

The LDA easily handles the case where the within-class frequencies are unequal and their performances has been examined on randomly generated test data as shown in Figure 5 where $m1$ and $m2$ are mean values. This method maximizes the ratio of between-class variance to the within-class variance in any particular data set thereby guaranteeing maximal separability. Then LDA is performed on the speaker models which are phonetically averaged from short segments and are reduced by the ANOVA pre-selection to estimate a reliable LDA transformation matrix.

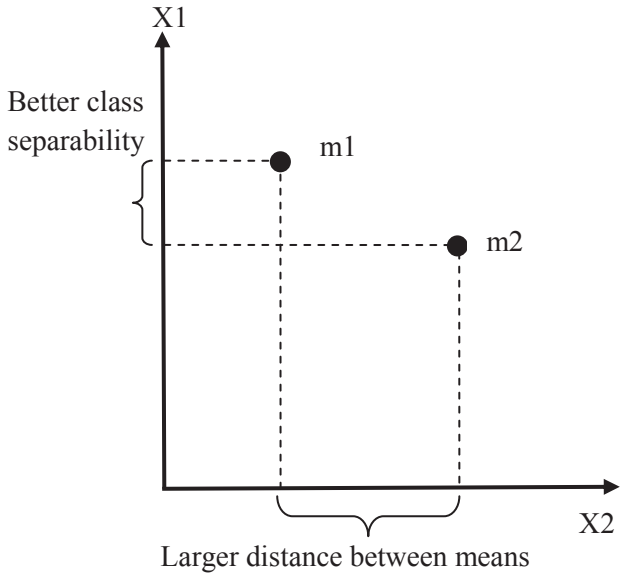

Fig. 5. LDA based class separation

D. Canonical Correlation Analysis (CCA)

The CCA techniques computes linear relationship between two multi dimensional variables say x and y. The two sets of basis vectors are found such that the correlation between the projections of the variables onto these basis vectors is maximized[24]. The Correlation Coefficients are determined as

$$S_x = xW_x \qquad S_y = yW_y \tag{8}$$

$$\rho = \frac{E\left[S_x, S_y\right]}{\sqrt{E\left[S_x^2\right] E\left[S_y^2\right]}} \tag{9}$$

IV. Conclusion

The field of speech processing consist of wide range of application areas that involve unique way of representing raw speech signal using features that not only reduce dimensionality issues but also storage issues. This work brings together the most popular feature extraction techniques spread across various domains and human biological models. The pros and cons of LPC, Complex Cepstrum, MFCC, HNM, aHM, Jitter and Shimmer are mentioned and studied in detail. Further, in systems involving larger databases and wide variety of features available for processing need selection algorithms that help in improving the performance of machine learning algorithm. Few feature selection techniques such as PCA, ANOVA, LDA and CCA are focused in this work.

References

[1] L. Rabiner and Biing-Hwang Juan,*Fundamentals of speech recognition*, 1993.

[2] Campbell Jr, P. Joseph, "Speaker recognition: A tutorial", Proceedings of the IEEE 85.9, pp. 1437-1462, 1997.

[3] D. Bees, B. Maier and P. Kabal, "Reverberant speech enhancement using cepstral processing", IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91, pp. 977-980, 1991.

[4] J. Nirmal, S. Patnaik, M. Zaveri and P. Kachare, "Complex cepstrum based voice conversion using radial basis function" ISRN Signal Processing, 2014.

[5] A. Nogueiras, A. Moreno, A. Bonafonte and J. Marino, "Speech emotion recognition using hidden Markov models", In INTERSPEECH, pp. 2679-2682. 2001.

[6] D. Mitrovic, M. Zeppelzauer and C. Breiteneder, "Features for content-based audio retrieval", Advances in computers 78, pp. 71-150, 2010.

[7] B. Atal and S. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", The Journal of the Acoustical Society of America 50.2B, pp. 637-655, 1971.

[8] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals", The Journal of the Acoustical Society of America 57.S1, pp. S35-S35, 1975.

[9] R. Vergin, D. O'shaughnessy and A. Farhat, "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition", IEEE Transactions on Speech and Audio Processing, vol. 7.5, pp. 525-532, 1999.

[10] A. N. Chadha, J. H. Nirmal and Pramod Kachare, "A Comparative Performance of Various Speech Analysis-Synthesis Techniques", International Journal of Signal Processing Systems 2.1, pp. 17-22, 2014.

[11] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis", IEEE Transaction on Speech Audio Processing, vol. 9.1, pp. 21-29, 2001.

[12] G. Degottex and Y. Stylianou, "A full-band adaptive harmonic representation of speech", In INTERSPEECH, 2012.

[13] X. Li, J. Tao, M. Johnson, J. Soltis, A. Savage, K. Leong and J. Newman, "Stress and emotion classification using jitter and shimmer features", IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 4, pp. IV-1081, 2007.

[14] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", The Journal of the Acoustical Society of America, vol. 87(4), pp.1738 - 1752, 1990.

[15] H. Hermansky, Hynek, N. Morgan, A. Bayya, and Phil Kohn "RASTA-PLP speech analysis technique", In IEEE Proceedings of International Conference on Acoustic, speech, and Signal Processing, pp. 121 - 124, 1992.

[16] D. Ververidis and C. Kotropoulos,"Emotional speech recognition: Resources, features, and methods", Speech communication, vol. 48(9), pp.1162-1181, 2006.

[17] J. Nirmal, M. Zaveri, S. Patnaik and P. Kachare, "Voice conversion system using salient sub-bands and radial basis function", Neural Computing and Applications, pp. 1 - 14, 2015.

[18] J. Nirmal, M. Zaveri, S. Patnaik and P. Kachare, “A novel voice conversion approach using admissible wavelet packet decomposition”, EURASIP Journal on Audio, Speech, and Music Processing, vol. 1, pp. 1 - 10, 2013.
[19] D. Koller and M. Sahami,, “Toward optimal feature selection”, Proceedings of International Conference on Machine Learning, 1996.
[20] M. Dash, H. Liu, “Feature Selection for Classification”, Journal of Intelligent Data Analysis, Elsevier, pp. 131 - 156, 1997.
[21] M. Sheikhan, M. Bejani and D. Gharavian, “Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method”, Neural Computing and Applications, vo. 23(1), pp. 215 - 227, 2013.
[22] H. Abdi and L. J. Williams, “Principal component analysis”, WIREs Computational Statistics John Wiley and Sons, Inc., vol. 2, 2010
[23] S. Balakrishnama and A. Ganapathiraju, “Linear discriminant analysis-a brief tutorial”, Institute for Signal and information Processing, 1998.
[24] D. R. Hardoon, S. Szedmak and J. S. Taylor, “Canonical correlation analysis: An overview with application to learning methods,Neural Computation”, vol. 16.12, pp. 2639-2664, 2004.