Analysis of Features from Analytic Representation of Speech using MP-ABX Measures

Pappagari Raghavendra Reddy, Karthika Vijayan, K. Sri Rama Murty

Department of Electrical Engineering
Indian Institute of Technology Hyderabad, India
{ee12m1023, ee11p011, ksrm}@iith.ac.in

Abstract

The significance of features derived from complex analytic domain representation of speech, for different applications, is investigated. Frequency domain linear prediction (FDLP) coefficients are derived from analytic magnitude and instantaneous frequency (IF) coefficients are derived from analytic phase of speech signals. Minimal pair ABX (MP-ABX) tasks are used to analyse different features and develop insights into the nature of information in them. The performance of the features derived from analytic representation are compared with performance of the Mel-Frequency Cepstral Coefficients (MFCC). It is noticed that the magnitude based features—FDLP and MFCC delivered promising PaC, PaT and CaT scores in MP-ABX tasks, demonstrating their phoneme discrimination abilities. Combining FDLP features with MFCC had proven beneficial in phoneme discrimination tasks. The IF features performed well in TaP mode of MP-ABX tasks, emphasizing the existence of speaker specific information in them. The IF significantly outperformed FDLP, MFCC and their combination in speaker discrimination task.

Index Terms: Analytic features, Instantaneous frequency, FDLP, MP-ABX tasks

1. Introduction

Feature extraction is a prominent task in speech processing and the capability of features to capture relevant information is a major factor deciding the efficiency of numerous speech processing algorithms. Hence, analysis of the information content in features becomes important. Evaluation of adeptness of features are usually carried out by examining their performances in recognition tasks. Recognition systems now a days employ some form of machine learning/modelling techniques. They may use supervised learning with labelled speech database, which is known to deliver faithful efficiency. In such cases, the effectiveness in terms of scores attained by the recognition system cannot be attributed to features alone. Hence this cannot be a legitimate procedure for evaluating features. We need much simpler strategies directly dealing with features themselves, rather than employing complex modelling processes.

The ABX measures were proposed as a suitable means for analysing features extracted from speech signals. ABX tasks consist of context or speaker dependant discrimination tasks between pairs of speech stimuli. These measures were introduced in speech processing to study auditory perception of different sounds by both ears using listening tests [1]. The ABX measures were also used to study effects of accents within language on perception by children with speech difficulties [2]. Objective evaluation of features using ABX measures was presented in [3, 4, 5] where dynamic time warping (DTW) was employed to calculate similarity between pairs of speech stimuli. In [3] similarity measures between features were evaluated for discriminating words. In [4], significance of each intermediate stage in the extraction of MFCC/PLP is analysed using ABX measures. The noise robustness of various features was studied using ABX measures in [5].

In this paper, we analyse the performance of features derived from complex analytic domain representation of speech using MP-ABX tasks. The FDLP coefficients extracted from analytic magnitude and IF features extracted from analytic phase are considered. These features are tested using ABX tasks set and their performances are compared with that of MFCC features. It is observed that the magnitude based features—MFCC and FDLP are demonstrating better phoneme discrimination capability irrespective of speaker variance than phase based features. Whereas, the IF features extracted from analytic phase presented better speaker discrimination ability irrespective of phoneme variance than FDLP. Also the existence of complementary information in these features is explored by feature concatenation and examining resultant ABX scores.

The rest of the paper is organized as follows: Section 2 explains the ABX measures used in this study. Section 3 discusses analytic features extracted from speech which are being studied in this paper. The evaluation of analytic features with respect to ABX tasks is presented in Section 4. A detailed discussion on the nature of information present in different features, effects of feature extraction parameters, existence of complementary information etc. are presented in Section 5. In Section 6, we summarize the results of analysis of features.

2. MP-ABX discrimination tasks

We have adopted the minimal pair ABX (MP-ABX) discrimination tasks explained in [4] for our study. Pattern discrimination using MP-ABX tasks employ three speech stimuli—A, B and X triplet, which are consonant-vowel (CV) pairs and measure the discrimination capability between such CV pairs. In each task A and B differ by one phoneme or speaker, and X is chosen such that it is close to A or B. We have considered four discrimination tasks, namely, Phoneme across Context (PaC), Phoneme across Talker (PaT), Context across Talker (CaT) and Talker across Phoneme (TaP). In PaC task, all the stimuli in triplet are spoken by the same speaker, in order to evaluate the robustness of features across context. A and B differ by one phoneme and test stimulus X is chosen to have one phoneme in common with either A or B. To measure the invariance of features across speakers, PaT task is chosen. In this task, A and B are spoken by same speaker, but X by a different speaker.
Table 1: Example triplets for MP-ABX tasks. SP stands for speaker.

<table>
<thead>
<tr>
<th>Task</th>
<th>A</th>
<th>B</th>
<th>X</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>PaC</td>
<td>/ba/ SP1</td>
<td>/ga/ SP1</td>
<td>/ga/ SP1</td>
<td>B</td>
</tr>
<tr>
<td>PaT</td>
<td>/ba/ SP1</td>
<td>/ga/ SP1</td>
<td>/ba/ SP2</td>
<td>A</td>
</tr>
<tr>
<td>TaP</td>
<td>/ba/ SP1</td>
<td>/ba/ SP2</td>
<td>/ga/ SP1</td>
<td>A</td>
</tr>
<tr>
<td>CaT</td>
<td>/ba/ SP1</td>
<td>/ga/ SP1</td>
<td>/ga/ SP2</td>
<td>B</td>
</tr>
</tbody>
</table>

These two tasks intend to show the effectiveness of features in conveying linguistic information in speech signals irrespective of speaker as well as context, demonstrating their efficiency for speech recognition applications. In TaP task, discrimination ability of features in distinguishing speakers is measured, examining their usefulness to speaker recognition applications. A and B are spoken by two different speakers, but have same phoneme sequence. X consist of a different phoneme sequence, but spoken by the same speaker as of either A or B [4].

In all the above tasks, either speaker or context is kept invariant, which is a controlled condition and cannot be demanded in real world scenarios. CaT task is proposed to evaluate the robustness of features across different contexts and different speaker identities. In this task, the triplet is chosen as in PaC task, but X is spoken by a different speaker. This task assumes significance in all speech recognition applications where features are required to be robust across speakers as well as context. The summary of all tasks utilized in this work is given in Table.1.

To evaluate each discrimination task, we follow the procedure described in [4]. The aim of evaluating each task is to find out the stimuli closest to test stimuli X. We compute dissimilarity between the pairs of phoneme sequences: A-X and B-X using DTW. Cosine similarity is used as distance metric in DTW which is recommended by [3]. An error will be identified when X shows more similarity to the stimuli which is not the correct answer. Mean error is computed by averaging the number of errors across speakers and number of trials using different triplets.

3. Analytic features of speech

The complex analytic domain representation of a continuous time signal $s(t)$ is given by [6]:

$$s_a(t) = s(t) + js_h(t)$$  \(1\)

where $s_h(t)$ is the Hilbert transform of the real signal $s(t)$, which is expressed as $s_h(t) = \mathcal{F}^{-1}\{S_h(j\Omega)\}$. $\mathcal{F}^{-1}$ denotes inverse Fourier transform and $S_h(j\Omega)$ is obtained as:

$$S_h(j\Omega) = \begin{cases} +jS(j\Omega) & \Omega < 0 \\ -jS(j\Omega) & \Omega > 0 \end{cases}$$  \(2\)

where $S(j\Omega)$ is the Fourier transform of $s(t)$ [7]. The analytic signal $s_a(t)$ can be expressed in polar form as

$$s_a(t) = a(t)e^{j\phi(t)}$$  \(3\)

where $a(t)$ and $\phi(t)$ are the time-varying magnitude and phase of the analytic signal, respectively. If $s(t)$ is a narrowband signal, $a(t)$ and $\phi(t)$ can be perceived as amplitude modulated (AM) and frequency modulated (FM) components of $s(t)$ [6].

Natural signals, including speech, cannot be represented by single time varying AM-FM component as they are mostly wideband signals. The multiband AM-FM demodulation of wideband speech signals was described in [8] and was utilized for formant tracking in [9, 10] using filterbank analysis. Features obtained from AM-FM demodulation were employed for speech recognition [11, 12], speaker recognition [13, 14, 15] etc. A narrowband component extracted from speech signal in Figure 1(a) is shown in Figure 1(b), together with its AM and FM components in Figure 1(c) and Figure 1(d) respectively.

3.1. Frequency domain linear prediction

Feature extraction from AM component using autoregressive modelling of temporal envelope of speech was discussed in [16, 17] which is the frequency domain dual of conventional linear prediction (LP) analysis. While LP analysis captures autocorrelations from time domain and models the magnitude spectral envelope in frequency domain, the FDLP analysis captures spectral autocorrelations and models temporal envelope. Analysis of temporal resolution of FDLP feature extraction was done in [18] and noise robustness capabilities were studies which showed significant improvements in phoneme recognition performance [19]. The temporal envelope obtained from a 16th order FDLP analysis of the signal in Figure 1(b) is shown in Figure 1(f), which demonstrates credible approximation of the original envelope of the signal in Figure 1(c).

3.2. Instantaneous frequency

The direct computation of analytic phase suffers from phase wrapping problem. It is not possible to draw meaningful inferences from analytic phase directly (see Figure 1(d)). But the
time derivative of analytic phase, the IF is free from phase wrapping and can be computed unambiguously. IF is the frequency of a sinusoid which locally fits a signal and wideband signals cannot be represented accurately with single sinusoid. Hence IF computation is meaningful only when the signal under consideration is decomposed into narrowband components. The IF computed from a narrowband component of speech shows variations in its frequency, around the centre frequency of corresponding narrowband filter (See Figure 1(e)). These variations correspond to formant transitions of speech. The post processing and extraction of features from IF is explained in [15] and are explored further using ABX tasks in this paper.

4. Evaluation of analytic features

The ABX triplets for MP-ABX tasks were obtained from TIMIT database by segregating required CV sequences from continuous speech. The phonetic level transcription available with the TIMIT database was used to extract these CV sequences. Each of the MP-ABX tasks was performed upon an approximate number of 5000 triplets spoken by 190 male and 90 female speakers.

The speech signals were sampled at a frequency of 16kHz. The required features - MFCC, FDLP and IF were extracted from 25 ms long segments of CV sequences shifted by 10 ms. Standard 39 dimensional MFCC were extracted using mel filterbank of length 47. Similarly, 39 dimensional FDLP features were extracted using a 47-filters mel filterbank. The length of filterbank used for IF feature extraction is studied further, as IF is strictly a property of narrowband signal, where filter bandwidths and spacing crucially affect its effective computation.

In each MP-ABX task, dissimilarity scores between A-X and B-X pairs were obtained by performing DTW on their features. The outcome of each trial was decided by comparing the dissimilarity scores from the A-X and B-X pairs. Mean error score (in %) was calculated by averaging the number of failed trials over the entire set of triplets under consideration.

Analysis of filterbank parameters in IF based feature extraction is done with respect to MP-ABX tasks. The mel spacing and linear spacing of filters in the filterbank is explored initially. The comparison between mel and linear filterbanks with different lengths, is shown in Figure 2. The performance of mel filterbank is consistently less than that of linear filterbank at all lengths. The frequency resolution of mel filterbank decreases considerably at high frequency (HF) ranges and thus the bandwidths of filters with higher centre frequencies are much high compared to those at low frequency range. Thus filtered components from HF region of speech signal lose their narrowband property, resulting in inefficient computation of IF and consequently poor performance. Hence linear filterbank is chosen for IF based feature extraction.

The effects of lengths of filterbank are studied using MP-ABX tasks, the results of which is shown in Figure 2. When the number of filters is very less, their bandwidths have to be increased to span the entire speech spectral range under consideration. Thus the filters will cease to be narrowband, making the IF computation inefficient, resulting in poor performance. On the other hand, when there exist exceedingly large number of filters in the bank, each filter will have minimal bandwidths insufficient to capture the frequency variations around its centre frequency. This will result in grave information loss upon IF computation, delivering increased errors. The length of filterbank is fixed as 37, based on the results of this analysis.

In [15], discrete cosine transform (DCT) is applied over the IF values computed from different frequency bands and first few DCT coefficients are retained, to account for redundancies within. The variation of mean errors with respect to number of DCT coefficients retained is shown in Figure 3. We have observed that the number of DCT coefficients to be retained should be greater than or equal to number of filters to obtain appropriate mean errors. This implies to the possibility of very less or no redundancy between IF computed from different frequency bands. As IF shows the local deviation in frequency about the centre frequency, the chances of existence of redundancies within information across neighbouring frequency bands is very less. As the extend of redundancy is very less, the DCT based compression results in information loss. Hence we have omitted the use of DCT and used the average values of IF over segments of speech from 37 frequency bands to obtain 37 dimensional feature vectors.

The MP-ABX tasks are performed on MFCC, FDLP and IF features and the mean errors for all tasks are shown in Table 2. It can be seen that MFCC and FDLP are performing almost equivalently in PaC and PaT tasks, showing their effectiveness in conveying linguistic information in speech signals, making them suitable for speech recognition applications. FDLP is performing exceptionally well in CaT task, implying that it is robust across different contexts irrespective of the speaker identity. On the other hand, IF features clearly outperformed the magnitude based features in TaP, which shows its stronghold in demonstrating speaker specific information in speech signals. Fusion experiments, combining pairs of features were carried out in order to explore the existence of complementary information within them. The combination of features was done by linear fusion of DTW similarity matrices. It can be seen from Table 2 that combination of features improved the performances in MP-ABX tasks by a remarkable margin, establishing the reciprocation of information between pairs of features. The mean errors computed here are inferior to those in [4] as the CV pairs are not isolated recordings, but obtained from continuous speech. This degradation in performance with stimuli obtained from continuous speech was already reported in [5].

5. Discussions

The tasks PaC and PaT represent phoneme discrimination capability of features regardless of context and speakers respectively. The CaT also denotes phoneme discrimination capability of features, in a more real world sense. On the other hand, TaP unveils speaker recognition capability of features. As it is already presented in Table 2, the magnitude based features show faithful performances in PaC and PaT tasks, claiming their suitability to speech recognition tasks. A closer inspection to the PaC scores from MFCC and FDLP reveals that MFCC is delivering lesser error than FDLP. As MFCC are short time spectral features, they are expected to be effective in identifying phonemes from short time frames as context will not play any significant

<table>
<thead>
<tr>
<th>Feature</th>
<th>PaC</th>
<th>PaT</th>
<th>CaT</th>
<th>TaP</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+IF</td>
<td>17.13</td>
<td>19.43</td>
<td>35.94</td>
<td>15.3</td>
</tr>
<tr>
<td>FDLP+IF</td>
<td>15.58</td>
<td>16.3</td>
<td>23.4</td>
<td>20.91</td>
</tr>
<tr>
<td>MFCC+FDLP</td>
<td>21.97</td>
<td>18.81</td>
<td>23.35</td>
<td>33.41</td>
</tr>
<tr>
<td>IF</td>
<td>20.99</td>
<td>25.95</td>
<td>39.48</td>
<td>15.78</td>
</tr>
<tr>
<td>MFCC</td>
<td>17.67</td>
<td>19.69</td>
<td>35.57</td>
<td>19.62</td>
</tr>
<tr>
<td>FDLP</td>
<td>23.35</td>
<td>23.35</td>
<td>33.41</td>
<td>33.41</td>
</tr>
<tr>
<td>IF</td>
<td>20.99</td>
<td>25.95</td>
<td>39.48</td>
<td>15.78</td>
</tr>
</tbody>
</table>

Table 2: Comparison of Mean Errors (%) for various features
role over short durations. But FDLP are extracted effectively from long time temporal envelopes, which can impart slight extend of context dependency into it and results in lesser performance than MFCC. In case of PaT task, both MFCC and FDLP are performing equivalently because speaker dependency will not get passed on to FDLP features as context dependency did. For CaT task, the scores attained from FDLP outperforms those from MFCC. FDLP succeeded in identifying phonemes irrespective of context and speaker, suggesting that speaker variance is more crucial than context variance in adversely affecting speech recognition. MFCC performed significantly superior to FDLP in TaP task, indicating its ability to convey relevant information for speaker recognition applications.

The IF features behaved almost complementary to FDLP in all tasks, demonstrating lesser context dependency and higher speaker dependency in phoneme recognition PaC and PaT tasks. Also, it is consistently poor in PaC, PaT and CaT tasks in comparison with MFCC. But, its speaker discrimination ability is distinctly visible from TaP task, where it delivered remarkable performance than MFCC and FDLP. This suggest that the magnitude characteristics of speech showcase pools of linguistic information, where as its phase characteristics convey mostly speaker specific information.

To analyse the presence of complementary information in features, fusion experiments were performed. The magnitude based features from Fourier domain and analytic domain were combined and MP-ABX tasks were carried out. It was observed that the short term spectral information and long time temporal information in MFCC and FDLP respectively, enacted effectively in bringing down the mean errors from their individual scores in all tasks regarding speech recognition. But the speaker discrimination ability of FDLP features is not evident, as it failed to improve the performance in TaP task, when combined with MFCC.

The FDLP and IF features were combined together to explore the magnitude and phase characteristics of speech in complex analytic domain. The complementary nature of information in analytic magnitude and phase as exhibited by FDLP and IF features succeeded in overcoming context and speaker dependencies in PaC and PaT tasks. The inability of IF features to dwindle down the adverse effects of context and speaker variance together in phoneme recognition made the performance of fusion suffer in CaT task. Similarly, the ineffectiveness of FDLP features in speaker discrimination obstructed the fusion from performing better in TaP task. Thus combination of MFCC and FDLP yielded the best performance in phoneme discrimination tasks and IF feature performed remarkably in speaker discrimination task.

6. Conclusions

Analysis of features of speech extracted from magnitude and phase of the complex analytic representation was carried out using MP-ABX tasks. The frequency domain linear prediction (FDLP) coefficients and instantaneous frequency (IF) features were obtained from analytic magnitude and analytic phase respectively. Their performances with respect to phoneme and speaker discriminative MP-ABX tasks based on CV pair speech stimuli were evaluated and compared with those of conventional MFCC features. It was observed that the FDLP features are efficient in conveying linguistic information in speech signals and upon combination with MFCC delivered exceptional phoneme discrimination performances. On the other hand, IF features had a stronghold in manifesting speaker specific characteristics of speech and outperformed MFCC, FDLP and their combination in speaker discrimination task. This study suggests the importance of magnitude of speech in conveying linguistic details and phase of speech in revealing speaker characteristics.
7. References


