Continuous Speech Recognition using Joint Features derived from The Modified Group Delay Function and MFCC

Rajesh M. Hegde, Hema A. Murthy

Department of Computer Science and Engineering
Indian Institute of Technology, Madras, Chennai.
{rajesh,hema}@lantana.tenet.res.in

Gadde V. Ramana Rao

STAR Laboratory, SRI International, 333, Ravenswood Avenue, Menlo Park, CA 94025
{rao}@speech.sri.com

Abstract
Feature extraction and selection for continuous speech recognition is a complex task. State of the art speech recognition systems use features that are derived by ignoring the Fourier transform phase. In our earlier studies we have shown the efficacy of The Modified Group Delay Feature (MODGDF) derived from the Fourier transform phase for phoneme, syllable and speaker recognition. In this paper we use the MODGDF and the popular MFCC derived from Fourier transform magnitude to compute joint features for continuous speech recognition of two Indian languages Tamil and Telugu. A novel method of segmentation of the continuous speech signal into syllable like units followed by isolated style recognition using HMMs is used. We further use an innovative technique which transforms the problem of detecting the correct string of syllabic units with maximum likelihood to finding an optimal state sequence locally. The recognition system does not use any language models. The MODGDF gave promising recognition performance for the two languages and compared well with the MFCC. Joint features derived using MODGDF and MFCC gave a 10.6% improvement for both Tamil and Telugu languages. The improvement reinforces the hypothesis that MODGDF captures complementary information to that of the MFCC and can be used along with the MFCC to capture the complete information in the speech signal at functional level and help in avoiding heavy auditory and language models.

1. Introduction
Contemporary speech signal analysis methods for continuous speech recognition are based on Fourier transform magnitude spectra. Features based on functional principles, like the LPCC, MFCC, PLP and RASTA-PLP have been used in this context. Auditory features simulating properties of the human auditory system have also been used, ignoring methods by which complete information about the speech signal can be derived from the functional principles alone. One such extremely important representation of the speech signal based on functional principles is the Fourier transform phase spectrum which has been almost ignored by speech researchers. We have proposed a new feature derived from the Fourier transform phase in our earlier efforts [1, 2, 3], called the modified group delay feature (MODGDF) and shown that it efficiently complements the MFCC derived from the Fourier transform magnitude for speaker and phoneme recognition tasks. In this effort we use the MODGDF and MFCC in isolation and also derive composite features by combining them to recognize Tamil and Telugu broadcast news. We use the concept of feature stream combination before the acoustic model [4] to combine the MODGDF and MFCC to derive the joint features. Traditional speech recognition systems use a very large vocabulary lexicon and heavy language models to achieve good recognition rates. The primary reason for this is that the baseline recognition techniques fail to identify or align the phonetic unit boundaries correctly. To overcome this problem large vocabulary lexicon and heavy language models are in place for building state of the art speech recognition systems. We therefore use a novel method of segmentation of continuous speech into syllable like units, using minimum phase group delay functions derived from the root cepstrum [5, 6], followed by isolated style recognition of these units. Once the recognition of a phrase of speech is over, an innovative method of identifying the correct syllable string by optimizing a set of possible syllable strings is used to improve recognition performance locally. This local optimization is extended to the whole phrase incrementally. The efficacy of the joint features is tested on two Indian (Tamil and Telugu) language databases [7] using the aforementioned baseline system. The recognition results are presented and its implications discussed in the final part of the paper.

2. The Modified Group Delay Function
Speech is completely characterized by both magnitude and phase information. The phase spectrum is available in a wrapped form which calls for unwrapping the phase spectrum to efficiently use it for any meaningful purposes. But the group delay function [3], defined as the negative derivative of phase, can be effectively used to extract various sys-
tem parameters when the signal under consideration is a minimum phase signal. The group delay function is defined as

\[ \tau(\omega) = -\frac{d(\theta(\omega))}{d\omega} \]  

where \( \theta(\omega) \) is the unwrapped phase function. The group delay function can also be computed from the speech signal as in [3] using

\[ \tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{[X(\omega)]^2} \]  

where the subscripts \( R \) and \( I \) denote the real and imaginary parts of the Fourier transform. \( X(\omega) \) and \( Y(\omega) \) are the Fourier transforms of \( x(n) \) and \( nx(n) \), respectively. The group delay function requires that the signal be minimum phase or that the poles of the transfer function be well within the unit circle for it to be well behaved. This has been clearly illustrated in [2] and [3]. It is also important to note that the denominator term \( |X(\omega)|^2 \) in equation 2 becomes zero, at zeros that are located close to the unit circle. The spiky nature of the group delay spectrum can be overcome by replacing the term \( |X(\omega)| \) in the denominator of the group delay function with its cepstrally smoothed version, \( S(\omega) \). Further it has been established in [1] that peaks at the formant locations are very spiky in nature. To reduce these spikes two new parameters \( \alpha \) and \( \gamma \) are introduced. The new modified group delay function as in [3] is defined as

\[ \tau_m(\omega) = \left( \frac{\tau(\omega)}{|\tau(\omega)|} \right) \left( |\tau(\omega)| \right)^{\alpha} \]  

where

\[ \tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}} \]  

where \( S(\omega) \) is the smoothed version of \( |X(\omega)| \). The new parameters \( \alpha \) and \( \gamma \) introduced vary from 0 to 1 where \( 0 < \alpha \leq 1.0 \) and \( 0 < \gamma \leq 1.0 \). The algorithm for computation of the modified group delay function is explicitly dealt with in [3].

2.1. Feature Extraction using the modified group delay function

To convert the modified group delay function to some meaningful parameters, the group delay function is converted to cepstra using the Discrete Cosine Transform (DCT).

\[ c(n) = \sum_{k=0}^{k=N_f} \tau_x(k) \cos(n(2k + 1)\pi/N_f) \]  

where \( N_f \) is the DFT order and \( \tau_x(k) \) is the group delay function. The second form of the DCT, DCT-II is used, which has asymptotic properties to that of the Karhunen Loeve Transformation (KLT) as in [3]. The DCT acts as a linear decorrelator, which allows the use of diagonal co-variances in modelling the speaker vector distribution.

3. Joint Features

Combining feature streams is a very ubiquitous and a flexible technique to improve speech recognition performance. Stream combinations before and after the acoustic model [4] are the most widely used methods although new techniques like the Tandem feature extraction have evolved. Review of literature on combining feature streams seems to suggest that the posterior combination of features is close to the formally correct approach when the streams are conditionally independent given the phonetic unit. But conditional independence of features given a phonetic unit is truly beyond the realm of human speech production. Since the intent is to capitalize on the practical differences between two features we strongly feel that combining streams at the feature level is the right approach given that the task is continuous speech recognition where each phonetic unit is not articulated in isolation. Hence we propose to use the following method to derive Joint features.

- Compute 13 dimensional MODGDF and the MFCC \(^1\) streams appended with velocity, acceleration and energy parameters.
- Use Feature stream combination to append the 42 dimensional MODGDF stream to the 42 dimensional MFCC stream to derive a 84 dimensional joint feature stream.

4. Performance Evaluation

In this section we discuss the corpora used, baseline and enhanced system to evaluate the individual and Joint features.

4.1. Corpora

Broadcast news corpora [7] of two Indian languages Tamil and Telugu are used to evaluate the performance. Both the corpora consist of speech files split into phrases, their corresponding sentence, word and syllable transcriptions.

- Tamil Corpora: This corpus consists of 20 news bulletins of Tamil language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 2184.
- Telugu Corpora: This corpus consists of 20 news bulletins of Telugu language transmitted by Doordarshan India, each of 15 minutes duration comprising 10 male and 10 female speakers. The total number of distinct syllables is 1896.

4.2. The Baseline recognition system

The baseline recognition system uses Hidden Markov Models trained apriori for 320 syllables for Tamil and 265 syllables for Telugu corpora [8].

\(^{1}\) Theory of MFCC is not dealt with in this paper as they are the features of choice in speech research.
bles for Telugu. The number of syllables used for training are selected based on their frequency of occurrence in the respective corpora. Any syllable that occurs more than 50 times in the corpus is selected as a candidate for which HMMs are built. A separate model is built for silence. It is significant to note that the reduction in vocabulary size for Tamil is 83.12% and Telugu is 87.87% using this scheme. Although the out of vocabulary syllable population is high in such a pruning scheme, the assumption is that the recognized syllable has to be phonetically nearer to one of the syllables in the vocabulary. The test phrase is segmented at boundaries of syllabic units using the minimum phase group delay function derived from the root cepstrum \[5, 6\]. These segments are now checked in isolated style against all HMMs built a priori using the reduced vocabulary. The HMM that gives the maximum likelihood value is declared as the correct match. Finally the output syllable sequence is generated by concatenating all the isolated syllables preserving the order of the original phrase. The recognition results using the baseline system for one news bulletin of duration 15 minutes comprising 4700 syllables for Tamil and Telugu are presented in Tables 1 and 2 respectively.

<table>
<thead>
<tr>
<th>Category</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole syllable</td>
<td>34.1%</td>
<td>32.1%</td>
<td>42.1%</td>
</tr>
<tr>
<td>Similar syllables</td>
<td>12.3%</td>
<td>11.7%</td>
<td>17.3%</td>
</tr>
<tr>
<td>Only vowel part</td>
<td>4.6%</td>
<td>4.1%</td>
<td>6.3%</td>
</tr>
<tr>
<td>Only consonant part</td>
<td>2.7%</td>
<td>2.9%</td>
<td>3.0%</td>
</tr>
<tr>
<td>Total</td>
<td>53.7%</td>
<td>50.8%</td>
<td>68.7%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole syllable</td>
<td>34.6%</td>
<td>32.6%</td>
<td>42.6%</td>
</tr>
<tr>
<td>Similar syllables</td>
<td>12.8%</td>
<td>12.2%</td>
<td>17.8%</td>
</tr>
<tr>
<td>Only vowel part</td>
<td>5.1%</td>
<td>4.6%</td>
<td>6.8%</td>
</tr>
<tr>
<td>Only consonant part</td>
<td>3.2%</td>
<td>3.4%</td>
<td>3.5%</td>
</tr>
<tr>
<td>Total</td>
<td>55.7%</td>
<td>52.8%</td>
<td>70.7%</td>
</tr>
</tbody>
</table>

### 4.3. The enhanced recognition system with local forced Viterbi realignment

To improve the performance of the baseline system the following approach is used:

- The HMM decoder is used to generate \( n \) best likely alternative syllabic units corresponding to each isolated test segment.
- All possible permutations with these \( n \) basic units at each position is enumerated and a list of possible syllable strings (sequences) is found. The correct syllable string (sequence) should be one among the list of possible syllable strings (sequences).
- The problem of finding the string of syllabic units with maximum likelihood is now transformed to the problem of finding an optimal state sequence.

To implement the enhanced system with local forced Viterbi realignment, we start with the complete recognized phrase output of the baseline system. Three consecutive syllables from the recognized phrase with ten possible alternatives to each syllable are considered. This now can be viewed as a concatenation of HMMs in parallel where each HMM corresponds to a syllable string of three syllables. This is equivalent to applying viterbi realignment locally over a set of three syllables. The realignment is done in sets of three syllables across the length of the phrase. A typical network for a particular syllable string /I rO ju/ for Telugu language consisting of three syllables is Figure 1. Note that a check mark is placed against the right syllable sequence in the network. The system does not use any heavy language models to rescore sentences according to grammar or semantics to address errors in segmentation. But it does rescore the recognized phrase coming from the baseline system by applying forced Viterbi realignment locally. This makes the system more implicit, less dependent on linguistic information and poses enormous new possibilities. The recognition results using the enhanced system for one news bulletin of duration fifteen minutes comprising 4700 syllables for Tamil and Telugu are illustrated in Tables 3 and 4 respectively.

<table>
<thead>
<tr>
<th>Category</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole syllable</td>
<td>37.1%</td>
<td>35.1%</td>
<td>48.9%</td>
</tr>
<tr>
<td>Similar syllables</td>
<td>13.3%</td>
<td>12.5%</td>
<td>23.3%</td>
</tr>
<tr>
<td>Only vowel part</td>
<td>5.3%</td>
<td>4.6%</td>
<td>9.3%</td>
</tr>
<tr>
<td>Only consonant part</td>
<td>2.9%</td>
<td>3.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Total</td>
<td>58.6%</td>
<td>55.2%</td>
<td>86.5%</td>
</tr>
</tbody>
</table>
Table 4: Recognition performance of the enhanced system for Telugu language (I) MFCC (II) MODGDF and (III) JOINT FEATURES

<table>
<thead>
<tr>
<th>Category</th>
<th>(I)</th>
<th>(II)</th>
<th>(III)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole syllable</td>
<td>39.6%</td>
<td>36.6%</td>
<td>50.6%</td>
</tr>
<tr>
<td>Similar syllables</td>
<td>14.8%</td>
<td>14.2%</td>
<td>22.8%</td>
</tr>
<tr>
<td>Only vowel part</td>
<td>6.1%</td>
<td>5.6%</td>
<td>8.8%</td>
</tr>
<tr>
<td>Only consonant part</td>
<td>3.2%</td>
<td>4.4%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Total</td>
<td>63.7%</td>
<td>60.8%</td>
<td>87.7%</td>
</tr>
</tbody>
</table>

present the results of one phrase from the Tamil news corpus where the baseline system wrongly identifies two syllables and the enhanced system is able to correct them by local forced Viterbi realignment.

Original syllable sequence:

muQn QnAL mu da la maic car sel vi je ya la li tA

Recognized syllable sequence of the baseline system:

muQn QnAL mu var lA mai ci su Lai vi je ya rar du tA SIL

Recognized syllable sequence after local forced viterbi realignment:

muQn QnAL mu da lA maic ci sel Lai vi je ya la du tA SIL

The syllables /da/, /maic/, /sel/ and /la/ which are wrongly identified by the baseline system are corrected by the enhanced system with local forced viterbi realignment.

5. Discussion and Conclusions

The reduced vocabulary size as a fraction of the original vocabulary size, as used in the baseline system is depicted as a pie chart in Figure 2 and the overall recognition percentage obtained using the MODGDF, MFCC and joint features is illustrated as a histogram in Figure 3.

The following conclusions can be drawn from the results presented in this paper:

- The new feature MODGDF derived from the phase spectrum performs on par with the highly popular MFCC and on combination with MFCC gives an overall improvement of 18% indicating that it captures complementary information to that of the MFCC.

- Use of joint features at feature extraction stage and the novel recognition system without the use of language models at the classification stage gives a good overall recognition performance of 86-87% and also reduces the original vocabulary size by 83-88%.

- The efficacy of the new feature MODGDF is proved for continuous speech recognition.

We intend to refine the MODGDF in terms of efficiently capturing the dynamic information of the speech signal in future. We also plan to propose certain improvements at the decoding stage of the enhanced recognition system to further improve recognition performance.

6. References


[4] D.Ellis, Feature stream combination before and/or after the acoustic model, ICSI Technical Report TR-00-007

