Tandem Processing of Fepstrum Features

Vivek Tyagi
IBM India Research Lab
New Delhi, India
vityagi1@in.ibm.com

Abstract

In our previous work [1, 2], we have introduced Fepstrum - an improved modulation spectrum estimation technique that overcomes certain theoretical as well as practical shortcomings in the previously published modulation spectrum related techniques[7, 8, 9]. In this paper, we provide further extensive ASR results using the Tandem processed Fepstrum features over the TIMIT corpus. The results are compared with TRAPS features derived from hierarchical and parallel structures of neural networks[3]. Unlike the multiple neural networks trained over multiple time-frequency patches or the frequency bands as in [3], we train a single neural network with the concatenated Fepstrum and MFCC features to derive Tandem(Fepstrum+MFCC) features. The resultant phoneme recognition accuracy of the concatenated Tandem(Fepstrum+MFCC)+MFCC feature is 76.5% on the TIMIT core test set and 77.6% on the complete test set making these one of the best reported results on the TIMIT continuous phoneme recognition task.

Index Terms: Fepstrum, Modulation spectrum, Tandem processing

1. Introduction

Several feature representations based on the broad concept of modulation spectrum have been proposed to describe the non-stationarity (spectral dynamics) inherent in the speech signal. Some representative examples are temporal patterns (TRAPS) features[5] and several modulation spectrum related techniques[7, 8, 9]. However, the notion of modulation spectrum in these techniques has been defined and implemented in a slightly ad-hoc manner. For instance, several researchers have extracted the speech modulation spectrum by computing a discrete Fourier transform (DFT) of the Mel or critical band spectral energy trajectories, where each sample of the trajectory has been obtained through a power spectrum (followed by Mel filtering) over 20-30 ms long windows. The major limitation of such a technique is that

- It implicitly assumes that within each Mel or critical band, the amplitude modulation (AM) signal remains constant within the duration of the window length that is typically 20-30 ms long.
- Instead of modeling the constantly and slowly changing amplitude modulation signal in each band, it mostly models the spurious and abrupt modulation frequency changes that occur due to the frame shifting of 10ms.

In our previous work[2], we have proposed a principled approach to estimate the amplitude modulation (AM) and the frequency modulation (FM) signals of the speech signal. The algorithm performs AM-FM demodulation of the speech signal in the time domain. As the AM-FM signal model is defined in the time domain\(^1\), a demodulation in the time domain leads to robust estimation of the continuously, though slowly evolving AM signals. It also leads to a better understanding of the relationships between various signal sub-components[2]. The lower DCT coefficients of these AM signals are then retained as the feature vector and is called Fepstrum. We have shown in [2] that Fepstrum is an exact signal processing dual of the well known quantity, real cepstrum. In [1], it was shown that a simple concatenation of the Fepstrum and MFCC features, trained over a hidden Markov model Gaussian mixture model (HMM-GMM) system, achieved a phoneme recognition accuracy of 74.6% on the TIMIT core-test, which is 1.8% better than the accuracy of the MFCC features (72.8%). In this work, we have used the Fepstrum features in conjunction with the Tandem processing[5, 4, 6]. In the next section we will provide our motivations to use the Tandem processing followed by the description of the system architecture and results.

2. Tandem processing

In this work, we have trained a single multi layer perceptron (MLP) with the concatenated Fepstrum and MFCC feature. The phone posterior probabilities at the output of the MLP are then again concatenated with the MFCC feature. This composite feature vector is then used to train the HMM-GMM system. This is illustrated as the System-A in Fig. 1. This processing has come to be known as the Tandem processing[5]. Here are some of the advantages of the tandem processing.

- The novel features may have varied physical meanings (speaking rate, articulatory features, modulation spectrum, etc.) and hence may have varied dynamic ranges, varied probability distributions and varied inter-dimensional correlations. Therefore, the direct HMM-GMM modeling of these features with diagonal covariances has not been found to be very effective. MLPs are effective at modeling unknown distributions[4]. Therefore they provide an effective interface to combine these novel features with the spectral energy based features such as MFCC as it transforms any arbitrary feature (Fepstrum, articulatory features, TRAPS) to the same physical quantity: the posterior probabilities\(^2\). These posterior probabilities may then be concatenated with the spectral energy features such as MFCC to train a HMM-GMM system[6, 4].

\(\text{\footnotesize\(x(t) = a(t) \cos(\int_0^t 2\pi f(u)du),\) here \(x(t)\) is a narrow band-pass filtered speech signal where, \(a(t)\) is the corresponding AM signal and \(f(u)\) is the corresponding FM signal.} \)

\(\text{\footnotesize\(^2\)In a way, MLPs can be seen as the "transistors" of the speech technology that perform logical combination of different features(input) streams.} \)
• As the MLP is trained on the concatenated Fepstrum and MFCC features, it learns the correlations between the two features and their joint distribution and outputs the posterior probabilities of the phones conditioned on the Fepstrum and the MFCC feature. This is particularly useful when the two features carry complementary information.

• The second pass of training (when the concatenated posteriors and MFCC feature are used to train the HMM-GMM system) can be seen as a non-linear error correction of the MLP posteriors. Suppose that, after training the MLP, it achieves a frame classification accuracy of 70%. A certain proportion of the 30% frames that are in error may have the "correct" phone in the second or third position. For example consider an incorrectly classified frame with the correct label /ae/ where the MLP’s posterior probabilities are: $P_{\text{mlp}}(q = /ae/) = 0.4$, $P_{\text{mlp}}(q = /ae/) = 0.3$ and $P_{\text{mlp}}(\text{rest of the phones}) = 0.3$. Here the correct label /ae/ is not so much further away from the incorrect label /a/. Now suppose that the MFCC feature vector for this frame is such that its HMM-GMM likelihood is higher for the correct phone /ae/ than the phone /a/. Then, the second pass of the HMM-GMM training can learn to correct this error as it also has the MFCC feature information.

2.1. Fepstrum Feature

In this section we briefly describe the AM signal extraction of the speech signal. A detailed description can be found in the previous work[1, 2]. We consider narrow band-pass filtered speech signal $x(t)$ and its analytic version $s(t)$ as

$$s(t) = x(t) + j\hat{x}(t)$$

where $\hat{x}(t)$ denotes the Hilbert transform of $x(t)$. In [2], we have shown that $s(t)$ can be factored as follows:

$$s(t) = A_\alpha e^{\alpha(t)} e^{\alpha\gamma(t)}$$

where $A_\alpha$ is a constant, $\alpha(t)$ is the logarithm of the AM signal, $\hat{\alpha}(t)$ its HT and $\hat{\alpha}(t) + \gamma(t)$ is the phase signal and its derivative is the FM signal.

From (2), we note that $\log|s(t)| = \alpha(t) + \log(A_\alpha)$, where $\log(A_\alpha)$ is a constant over the frame. Therefore the logarithm of the absolute magnitude of the analytic signal in each band is an estimate of the corresponding AM signal + a constant term. This is illustrated in Fig. 2. The obtained AM signal is a low frequency signal and hence is down-sampled. Finally its lower DCT coefficients that typically correspond to $[0, 25Hz]$ are retained as the feature vector: Fepstrum.

3. Experiments

We have performed the experiments on the TIMIT corpus, which has a bandwidth of 8KHz. Conventional twenty four Mel filters were used to decompose the wide-band speech analytic signal into narrow band-pass filtered analytic signals as in Fig. 2. The AM signal is obtained as the logarithm of the absolute magnitude of the narrow-band filter output. At this stage, the AM signal has the same sampling frequency as the original speech signal ($16KHz$). The AM signals are low modulation frequency signals[2]. Therefore, we filter the AM signals through a low-pass filter of cutoff-frequency 100 Hz and then downsample them by a factor of 80. Long rectangular windows of size 100 ms were used to frame the narrow-band pass filtered analytic signals. We chose a rectangular shape of the window to avoid any artificial tilt in the lower DCT coefficients. We then retain its first 5 DCT coefficients (Fepstrum) that correspond to $[0, 25 Hz]$. Fepstrum sub-vectors from each band are concatenated together to form a vector of dimensionality 120 ($5 \times 24$). We perform a KL transform (PCA) on this vector, followed by dimensionality reduction to obtain a 60 dimensional feature vector. This is illustrated in the Fig. 2.

3.1. Tandem-Fepstrum Training

Two Tandem feature sets were computed for comparing the ASR accuracy on the TIMIT core-test set and the complete test-set. Quicknet tools\(^4\) employing three layer perceptron with soft-max non-linearity at the output, were used for all the experiments.

1. Concat.Tandem(MFCC)+MFCC. A single MLP was trained on MFCC features. The input layer had $39 \times 9 = 351$ nodes (corresponding to the 9 frames of MFCC), 800 hidden nodes and 48 output nodes (corresponding to the 48 labels in the TIMIT corpus). Decorrelated phone posteriors at the output of the MLP were then concatenated with the raw MFCC feature vector that corresponded to the center ($5^{th}$ frame of the 9 frame input to the MLP. This composite feature vector was then used to train the HMM-GMM system. In [6], this composite feature provided significant recognition accuracy improvement on the Switchboard corpus. The choice of 9 frames and 800 hidden units provided the best recognition accuracy. Increasing the number of hidden nodes deteriorated the recognition performance. This is the system B in the Fig. 1 and it forms our Tandem-MFCC baseline.

2. Concat.Tandem(Fepstrum+MFCC)+MFCC. 60 dimensional Fepstrum was concatenated with the 39 dimensional MFCC to train a single MLP with $(60 + 39) \times 9 = 891$ input nodes, 800 hidden nodes and 48 output nodes. Again the phone posteriors were concatenated with the raw MFCC feature to train the HMM-GMM system. This is the system A in Fig.1. We chose to train the MLP with the concatenated Fepstrum+MFCC features so that the MLP could learn the correlations between the two features. The values of the time-context (9 frames) and the hidden nodes (800) were the same as in the above baseline system.

Following the convention, phonetic recognition accuracies are tabulated using the commonly adopted 39 labels (phones) after label folding and the results are provided in Tab. 1. Decision tree clustered triphone HMMs with 3 states per triphone were trained for each of the features. Finally each triphone state emission density was modeled with 11 component diagonal covariance Gaussians. A phone bi-gram language model was used for all the features and its weight was empirically optimized to achieve the best phoneme recognition accuracy\(^5\) on the baseline Concat.Tandem(MFCC)+MFCC system (LM weight=4). The same LM weights was used for the Fepstrum feature based systems as well.

In Tab. 2 we provide the results using only the HMM-GMM system. Upon comparing the two tables, we notice that the

\(^4\)Part of the SPRACH core package developed at ICSI, http://www.isi.edu/dpwe/projects/sprach

\(^5\)Including deletion, substitution and insertion errors
System A: Concat_MFCC_Tandem(Fepstrum_and_MFCC)

System B: Concat_MFCC_Tandem(MFCC)

Figure 1: Tandem processing of the Fepstrum and MFCC features.

Figure 2: The FEPSTRUM feature extraction

Table 1: Phoneme Recognition accuracies of the Tandem processed MFCC and Fepstrum. 2nd column: Core-test set, 3rd column: Complete test-set.

<table>
<thead>
<tr>
<th>System</th>
<th>Core-test set</th>
<th>Complete test-set</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>75.6</td>
<td>76.6</td>
</tr>
<tr>
<td>Concat_Tandem(MFCC)+MFCC</td>
<td>76.6</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Table 2: Phoneme recognition accuracy using only the HMM-GMM system on the TIMIT core-test set[1]

Table 3: Phoneme Recognition accuracies of hierarchical MLPs based TRAPS and Split-time context[3]. 2nd column: Core-test set, 3rd column: Complete test-set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Concat Fepstrum + MFCC</td>
<td>74.6</td>
<td>76.6</td>
</tr>
</tbody>
</table>

Concat.Tandem(MFCC)+MFCC i.e. the System-B has an accuracy improvement of (75.6 – 72.8 = 2.8%) over the MFCC based HMM-GMM baseline. This improvement can be attributed to the *error correction* mechanism of the Tandem processing as explained in the previous section. The Concat.Tandem(Fepstrum+MFCC)+MFCC has a further improved phoneme recognition accuracy at 76.6% on the core-test set and 77.6 on the complete test set.

In [3], the authors have investigated the TRAPS[5]; a hierarchical structure of several MLPs with separate classification of input patterns in critical bands and split temporal context (STC) systems[3]. They reported the best accuracies with 5 band TRAPS (with 6 MLPs) and 5 block split-time-context (with 6 MLPs) and the results are reproduced in Tab. 3. They trained the MLPs with 39 × 3 = 117 output nodes to get the posteriors for each of the three states of a phone. They showed that it improved the accuracy as compared to the case when just one output node is used per phone (as is the case in this work).

Despite these differences, we note that the single MLP based Concat. Tandem(Fepstrum+MFCC)+MFCC compares favorably with respect to these results. In [3], the authors have mentioned that they were able to further fine tune the 5 block STC system by fine tuning the MLP training. They added CV data to have more training data and modified the scheduler to train up to 20 epochs amongst other steps. These steps were not performed in this work.
4. Summary

In this paper we have compared the Tandem(Fepstrum+MFCC)+MFCC feature with the TRAPS and split-time-context features that are derived from a hierarchical structure of several MLPs operating on several critical bands and split-time contexts[3]. The Tandem(Fepstrum+MFCC)+MFCC feature used a single MLP and achieved results comparable to other TRAPS and STC systems[3].

While Fepstrum provides amplitude modulations (AM) occurring within a single frame of size 100ms, the MFCC feature provides a short-time spectrum over a 20ms window. Together these two features complement each other and their combination through simple concatenation offers improvements both at the HMM-GMM and the Tandem modeling levels.

5. References