Combining feature space discriminative training with long-term spectro-temporal features for noise-robust speech recognition

Takashi Fukuda, Osamu Ichikawa, and Masafumi Nishimura

IBM Research – Tokyo, IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi, Kanagawa, 242-8502, JAPAN
{fukuda1, ichikakw, nisimura}@jp.ibm.com

Abstract

Discriminative training of feature space using maximum mutual information (fMMI) objective function has been shown to yield remarkable accuracy improvements. For noisy environments, fMMI can be regarded as an effective noise compensation algorithm and can play a significant role for noise robustness. Feature space speaker adaptation techniques such as feature space maximum likelihood linear regression (fMLLR) are also well known, suitable for mismatched test data. These feature space transform algorithms are essential for modern speech recognition but still need further improvement against low SNR conditions. In contrast, long-term spectro-temporal information has also received attention to support traditional short-term features. We previously proposed long-term temporal features to improve ASR accuracy for low SNR speech. In this paper, we show that long-term temporal features can be combined with fMMI to build more discriminative models for noisy speech and the proposed method performed favorably at low SNR conditions.

Index Terms: Automatic speech recognition, feature extraction, feature space discriminative training, feature space speaker adaptation, long-term spectro-temporal feature

1. Introduction

To improve the robustness of automatic speech recognition (ASR), a lot of feature-based approaches such as PLP [1] and RASTA [2] have been proposed over the years. Among the methods, discriminative training of feature spaces using maximum mutual information (fMMI) or minimum phone error (fMPE) objective functions was recently shown to yield remarkable accuracy improvements [3, 4, 5]. The first stage of feature space discriminative training is to transform the features into a high dimensional space with a set of Gaussians created by likelihood-based clustering of the Gaussians in the acoustic model. The high dimensional features are projected down to the dimension of the original features and added to them. For noisy environments, fMMI can be regarded as an effective noise compensation algorithm. This is somewhat expected, given its close connection to the MMI-SPLICE algorithm [6].

Feature space maximum likelihood linear regression (fMLLR) and feature space maximum a posteriori linear regression (fMAPLR) are also well known algorithms for effective speaker adaptation, especially on mismatched test data [7, 8]. Although the feature space adaptation algorithms were originally proposed to compensate for channel and speaker variations, they were shown to work well also for noisy environments. Feature space discriminative training and speaker adaptation are essential for modern speech recognition, but there is still room for improvement against very low SNR conditions such as 5 dB or 0 dB in which we often encounter in vehicular environments.

Meanwhile long-term spectro-temporal information in speech has been used for improving noise robustness of ASR. Hermansky et al. proposed a method to represent information spanning more than 500 ms by using multi-layered perceptrons (MLPs) [9]. Their approach uses temporal patterns consisting of consecutive frames of log critical band energies and feeds these patterns into the MLPs. There are many related works in this area [10, 11]. In contrast, psychological tests by Poeppel showed that human beings use two types of temporal information extracted from both short (20 to 40 ms) and long (150 to 250 ms) temporal windows to understand spoken language [12]. These prior studies suggest that long-term temporal information has value for accurate ASR systems. In relation to these studies, we previously investigated the long-term delta features obtained by enlarging a delta window and found that combining these features with conventional short-term features improved the performance [13].

In this paper, we describe how feature space transforms can be made more noise robust by combining them with long-term dynamic features. Our goal in this paper is to further improve performance of fMMI and fMLLR to operate reliably down to very low SNR situations of 0 dB or even −5 dB without increasing large computational complexity.

This paper is organized as follows. Section 2 shows fundamentals of feature space transforms. Section 3 provides the effectiveness of the dynamic features as a long-term temporal information representation. Sections 4 cover the experiments. Finally, Section 5 finishes with our conclusions.

2. fMMI and fMLLR implementation

2.1. fMMI and fMLLR pipeline

The fMMI process can be described by two fundamental stages [5]. The first stage, level 1, relies on a set of Gaussians $G$ to convert an input $d$-dimensional feature vector $x_t$ to offset features

$$o(t, g, i) = \begin{cases} \gamma_g \frac{o^{(t)}}{\sigma^{(t)}}, & \text{if } i \leq d \\ \frac{5\gamma_g}{o^d}, & \text{if } i = d + 1, \end{cases}$$

where $t$ denotes time, and $i$ denotes offset dimension. $\gamma_g$ is the posterior probability of $g \in G$ given $x_t$. The set $G$, of size $G$, is arrived at by clustering the Gaussians of the original acoustic model.

In general $o(t, g, i)$ contains $(d+1)G$ elements for each time $t$. For computational efficiency all $\gamma_g$ below a threshold $\gamma_{cut}$ are set to 0 resulting in a sparse $o(t, g, i)$. The offset features are operated on by a level 1 transform $M_1(g, i, j, k)$

$$b(t, j, k) = \sum_{i=0}^{d} M_1(g, i, j, k) o(t, g, i) = \sum_{g \gamma_g \geq \gamma_{cut}} \sum_{i=0}^{d} M_1(g, i, j, k) o(t, g, i),$$

where $M_1$ is parameterized by Gaussian $g \in G$, offset dimension.
where $i \in \{1, \ldots, d+1\}$, an outer-context $j \in \{1, \ldots, 2J+1\}$ and final output dimension $k \in \{1 \ldots d\}$.

The next stage of fMMI, level 2, takes as input $b(t + \tau, j, k)$ for $\tau \in \{-\Lambda, \ldots, \Lambda\}$. It computes its output as

$$\delta(t, k) = \sum_j \sum_{\tau} M_2(j, k, \tau + \Lambda + 1) b(t + \tau, j, k),$$  

(3)

where $\tau$ is the inner context. The output of level 2, $\delta(t, k)$, is added to $x_i(k)$ to compute the fMMI features. The level 1 and 2 transforms are estimated in the training stage.

### 2.2. fMLLR and processing pipeline

The fMLLR algorithm was proposed in [7]. The feature vector $x_t$ is transformed as

$$\tilde{x}_t = Ax_t + b = W \xi_t,$$

(4)

where $A$ is the $n \times n$ rotation matrix, $b$ is the $n \times 1$ bias term, $\xi_t = [1 \ x_t^T]^T$ is the $(n + 1) \times 1$ extended observation vector and $W = [b \ A]$ is the $n \times (n + 1)$ extended transformation matrix. The transform parameters are estimated by optimizing the following auxiliary Q-function,

$$Q_{ML} = -\frac{1}{2} \sum_{t,m} \gamma_m(t) | \log |A| |^2$$

$$+ (W \xi_t - \mu_m)^T \Sigma_m^{-1}(W \xi_t - \mu_m),$$

(5)

where $\mu_m$ and $\Sigma_m$ are the mean and covariance for Gaussian component $m$ and $\gamma_m(t)$ is the posterior probability of being in Gaussian $m$ at time $t$.

### 3. Long-term dynamic features

#### 3.1. Dynamic feature extraction as a filtering process

Our previously proposed method incorporates long-term temporal information (long deltas) into a feature parameter set by combining conventional dynamic features extracted from both short- and long-term cepstrum sequences with linear regression calculation [13]. Here we discuss the long-term dynamic feature extraction in terms of a filtering process for a modulation spectrum. A number of studies of the modulation spectra have been done to assess the human phonetic perception. Drullman et al. showed that high-pass filtering above 4 Hz or low-pass filtering below 16 Hz for the modulation spectrum did not reduce the intelligibility of speech [14, 15]. Kanedera et al. found that most of the useful linguistic information in speech lies in the modulation frequencies ranging from 1 to 16 Hz and especially from 2 to 10 Hz [16].

A linear regression calculation used in the delta feature extraction is also regarded as a filtering process that emphasizes an essential component of the speech. In contrast, the long-term delta features focus on the slowly changing spectral variations by transforming the modulation spectrum. For the important modulation frequencies for ASR, such as around 10 Hz, it is meaningful that the long-term delta features handle modulation spectra of interest entirely from 2 to 10 Hz. This also corresponds with the results of the psychological tests by Poeppel [12]. In addition, the long-term delta features show significant improvement for noise-robust voice activity detection (VAD) by enabling the delta window beyond the average phoneme duration of an utterance [17].

#### 3.2. Long-term features for fMMI and fMLLR

The fMMI and fMLLR algorithms map the feature vector $x_t$ into a more canonical feature space by using a transformation matrix. There are several factors that influence the ASR performance when fMMI and fMLLR are used. A typical example in fMLLR is that when adaptation data is limited, not only the linear transform can be easily over-trained, but performance may not improve. The fMMI will be influenced by training data, learning rate, and a pair of inner and outer context. Furthermore, characteristics of feature vector $x_t$ input into fMMI and fMLLR are also important factors that affect ASR accuracy.

Here we discuss how temporal variations of spectra are considered in fMMI and fMLLR. By using fMLLR, it is only necessary to apply a linear transform to the feature vectors for every frame. The conventional fMLLR doesn’t consider adjacent frames but the current frame only. Therefore the feature vector $x_t$ for fMLLR should include long-term spectro-temporal information, otherwise there are no more opportunities to grasp temporal variations. In contrast, fMMI can deal with temporal variations by adjusting inner and outer contexts for level 1 and 2 transform. However, the input feature vector $x_t$ for fMMI is first mapped into a high dimensional feature space composed of posterior probabilities and adjacent frames are considered in the space. Hence temporal variations of posterior vectors in fMMI does not necessarily reflect spectro-temporal characteristics on human phonetic perceptions, and thus long-term spectro-temporal variations should also be included in the feature vector $x_t$. In our modulation frequencies around 10 Hz are emphasized by the short-term linear regression filtering with 7 frames while those around 2 Hz are emphasized by long-term linear regression filtering with 17 frames. The short-term delta features enhance the linguistic information with preserving the intelligibility of the speech. In contrast, the long-term delta features focus on the slowly changing spectral variations by transforming the modulation spectrum.
method, we use a combination of short and long delta features as an input to feature space transforms of fMMI and fMLLR.

4. Experimental setup

Two types of embedded ASR systems including medium-resource ASR and very light-resource ASR that needs quick response were tested here.

4.1. ASR testing with fMLLR + Long deltas

4.1.1. Speech data

An in-house English in-car speech corpus was used in our experiments for the medium-resource ASR. The recognition tasks involve an address, connected digit strings, an audio control command, places of interest, a free form command, and an isolated spoken term recorded in automotive environments at speeds of 0 mph, 30 mph, and 60 mph with a 16 kHz sampling frequency. There are approximately 796.1 hours of data in the training set, and 39.2 hours of data consisting of 38,905 sentences uttered by 133 speakers in the test set.

4.1.2. Features for fMLLR

The input speech is represented by a static feature vector of 13 dimensional MFCCs that are computed from 25-ms frames with a 15-ms shift. We compared three dynamic features for fMLLR. LDA: The LDA features are computed by splicing nine consecutive frames into a 117 dimensional supervector, then projecting it to 40 dimensions.

Short delta & delta-delta: Conventional short delta and delta-delta (referred to as ddelta in the experiments) features are extracted from the 13-dimension MFCCs. The window sizes for short delta and delta-delta features are 5 frames (75 ms) and 9 frames (135 ms), respectively. The total number of dimensions of the feature parameter set is 39.

Short & Long delta: The short delta features are extracted as already described. In contrast, the delta-delta feature is eliminated from the feature parameter set, and instead, the long-term delta features are applied to the feature parameter. Here we use long delta features only for the c4 – c12 coefficients of MFCC and the normal delta-delta features are applied to the c0 – c3 coefficients, considering the characteristic that the lower cepstra usually change slowly. The long delta feature is extracted with 11 frames (165 ms) in the experiment with a 15-ms frame shift. In another experiment using a 10-ms frame shift, we used 17 frames (170 ms) as the long delta features [13]. There are 39 dimensions in the feature parameter set including the 13-dimension static MFCC, their short delta, the short delta-deltas for c0 – c3, and the long delta for c4 – c12 coefficients.

The range of these three transformations was approximately diagonalized using a global semi-tied covariance (STC) matrix [18]. STC updates are interleaved with standard HMM updates during the maximum likelihood training of the acoustic model. Prior to splicing, projection, and delta feature calculation, the MFCC features are mean normalized on a per utterance basis.

4.1.3. Acoustic modeling

Words in the recognition lexicon are represented as sequences of phones, and the phones are modeled with 3-state left-to-right HMMs that do not permit state skipping. Acoustically distinct variants of HMM states are identified using decision trees that ask questions about the phonetic context in which a state occurs, and the leaves of the decision tree are the basic acoustic units that we model. All models have quinphone acoustic context. There were 830 context-dependent states corresponding to the leaves of the decision tree and approximately 10,000 Gaussian mixture components used in our experiments. We framed the decoding task as a search on a finite-state machine created by the offline composition of several finite-state transducers.

4.1.4. Experimental results

We used fMLLR per speaker to map the delta and LDA-based features to the canonical feature space. Figures 2 and 3 show the experimental results. The line chart shows the word error rate and the bar chart shows the number of sentences tested in each SNR bin. In automotive environments, we often have problems with the ASR performance in low SNR conditions including 5 dB or 0 dB rather than high SNR conditions. We focus on the word error rate below 10 dB here. Comparing Figures 2 and 3, fMLLR improved performance in every SNR bin regardless of the input feature vector for fMLLR. The fMLLR combined with the long delta features significantly reduced the word error rate as compared with fMLLR combined with LDA features and short delta-delta (ddelta) features especially below 10 dB. In another experiment using LDA with a larger context, the performance was worse than 9-frame-LDA. There are many cases in conventional methods that if fMLLR is used to map the features into a new space, gains on the feature space before doing fMLLR disappear in the fMLLR feature space. In contrast, the proposed feature set of fMLLR combined with long delta features steadily showed improvements.
Table 1: Performance of fMMI transform combined with long deltas.

<table>
<thead>
<tr>
<th>SNR bin number</th>
<th>w/o fMMI transform</th>
<th>w/ fMMI transform</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bin 0</td>
<td>bin 1</td>
</tr>
<tr>
<td>SNR center bin [dB]</td>
<td>1.0</td>
<td>9.5</td>
</tr>
<tr>
<td>Short delta &amp; ddelta (delta-delta)</td>
<td>10.6</td>
<td>5.1</td>
</tr>
<tr>
<td>Short delta &amp; Long delta</td>
<td>9.8</td>
<td>4.8</td>
</tr>
</tbody>
</table>

6. References