Multi-stream acoustic model adaptation for noisy speech recognition

Satoshi Tamura* and Satoru Hayamizu*
* Department of Information Science, Gifu University, Gifu, Japan
E-mail: tamura@info.gifu-u.ac.jp, hayamizu@gifu-u.ac.jp

***Abstract*—In this paper, a multi-stream-based model adaptation method is proposed for speech recognition in noisy or real environments. The proposed scheme comes from our experience about audio-visual model adaptation. At first, an acoustic feature vector is divided into several vectors (e.g. static, first-order and second-order dynamic vectors), namely streams. While adaptation, a stream performing relatively high recognition performance is updated for the stream only. Alternatively, a stream having less recognition power is adapted using all the streams that are superior to the stream. In order to evaluate the proposed technique, recognition experiments were conducted using every streams, and then adaptation experiments were also investigated for various types of combination of streams.**

I. Introduction

In order to overcome degradation of accuracy of Automatic Speech Recognition (ASR), a lot of approaches have been explored: signal enhancement such as beam forming, signal and feature compensation like Spectrum Subtraction (SS), acoustic model adaptation and audio-visual speech recognition. Acoustic model adaptation, that converts model parameters so as to match existing noises, is an essential technique for state-of-the-art ASR systems particularly in real environments. Most adaptation techniques are categorized into posteriori-probability-based methods (e.g.[1]), transformation-based schemes (e.g.[2]), and database-based strategies (e.g.[3]). This paper focuses on the former two approaches.

There is a remarkable research where static and dynamic features were compared and it turned out that dynamic information is more effective in noisy circumstances [4]. The authors further investigated a multi-stream recognizer that had static and dynamic streams, then achieved the improvement by controlling stream weights. This research indicates a possibility to improve recognition performance by treating an acoustic feature as a combination of multiple streams, like audio and visual streams in audio-visual speech recognition.

We have developed multi-modal (audio-visual) ASR which uses speech data and visual data, i.e. mouth/lip images, and model adaptation for audio-visual speech recognition was investigated [5]. In general, the performance of acoustic stream is better than that of visual stream except in heavily noisy conditions. Then we proposed an adaptation technique for audio-visual model, where audio model parameters are updated using only acoustic features whereas visual model parameters are adapted using audio-visual features. This scheme comes from the idea that the superior modality helps the inferior modality in the multi-modal adaptation, and improves the whole performance of multi-modal ASR. Since the conventional acoustic feature consists of static and dynamic parameters that have different properties, the same adaptation scheme might be applicable to audio-only ASR.

Inspired these researches, in this paper we propose a novel adaptation method derived from a multi-stream technique; an acoustic feature space is divided into several subspaces, namely streams. Pre-recognition is performed in each stream to evaluate the reliability. Model adaptation for a stream is subsequently conducted using the streams that have better performance than the current stream, before adapted models are integrated. Recognition results are finally generated by the adapted acoustic model. The improvement of recognition accuracy is expected by exploiting superior streams and ignoring inferior streams in model adaptation. Through an evaluation corpus, the following recognition accuracies were estimated and compared: original feature vector and every streams, those after model adaptation, and our proposed adaptation strategy.

The rest of this paper is organized as follows: Section II introduces conventional model adaptation methods as well as our proposed model adaptation technique. Recognition experiment and adaptation evaluation are presented in Section III. Section IV concludes this paper.

II. Model adaptation

A. Conventional methods

1) Maximum A Posteriori (MAP):

Maximum A Posteriori (MAP) is a method to re-estimate model parameters using adaptation data. Let us denote a model parameter set by θ, and a prior distribution of θ by $g(\theta)$. Now assume a feature vector $\boldsymbol{x}$ is observed and denote its probability by $p(\boldsymbol{x}|\theta)$. A posteriori probability of θ is then described as:

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)g(\theta)}{\int p(\boldsymbol{x}|\vartheta)g(\vartheta)d\vartheta} \tag{1}$$

MAP adaptation determines the model parameter so as to maximize the posteriori probability as:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(\boldsymbol{x}|\theta)g(\theta) \tag{2}$$

MAP can achieve better performance than MLLR described below, if enough adaptation data are available.

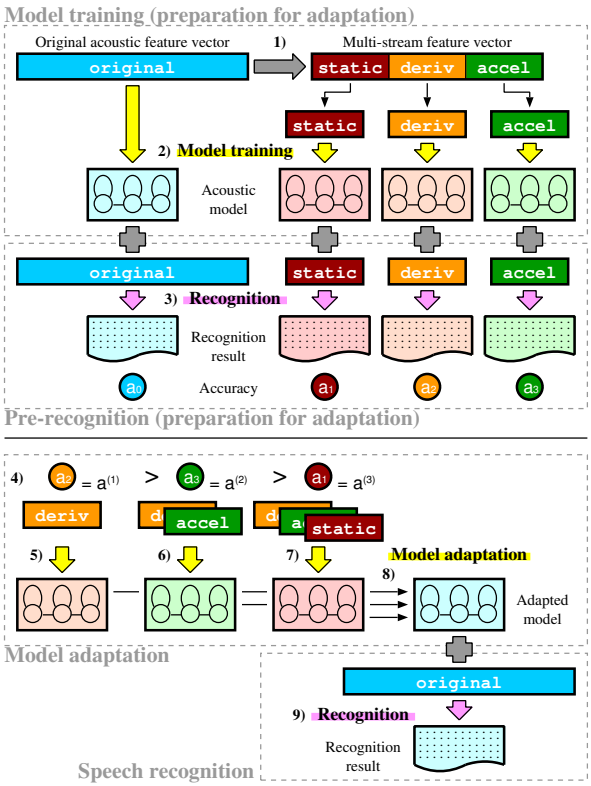

Fig. 1. A summary of proposed multi-stream model adaptation technique.

2) Maximum Likelihood Linear Regression (MLLR):

Maximum Likelihood Linear Regression (MLLR) [2] has been widely known in speech recognition. MLLR is typically used so as to adapt model parameters in noisy environments. Let us denote a D-dimensional mean vector of Gaussian distribution at a state in a Hidden Markov Model (HMM) by $\boldsymbol{\mu}$. MLLR projects the mean vector into an adapted vector $\hat{\boldsymbol{\mu}}$ by the following linear regression:

$$\hat{\boldsymbol{\mu}} = A\boldsymbol{\mu} + \boldsymbol{b} \tag{3}$$

where A is a D-dimensional square matrix and $\boldsymbol{b}$ is a bias vector. The equation (3) can be rewritten as:

$$\hat{\boldsymbol{\mu}} = W\boldsymbol{\xi} \tag{4}$$

where $\boldsymbol{\xi} = (1\ \boldsymbol{\mu}^\top)^\top$ and $W = (\boldsymbol{b}\ W)$. The matrix W can be obtained using adaptation data. Compared with MAP, MLLR has an advantage that it requires relatively less adaptation data.

B. Proposed adaptation method

We have investigated audio-visual interaction in model adaptation for audio-visual ASR (AVASR) [5]; how audio information affects visual adaptation, or how visual features contribute audio model adaptation. The conclusion in the work is that the superior modality (usually audio stream) helps the adaptation of the inferior modality (visual stream) since the superior stream have still significant information to discriminate compared with another stream. Consequently the performance of AVASR can be improved.

Motivated by the previous research, in this paper we propose a multi-stream adaptation technique for conventional audio-only ASR. Figure 1 illustrates the procedure of the proposed method, that is also explained in detail as follows:

1) Divide an input feature vector $\boldsymbol{x}_t$ into N partial vectors (called streams) $\boldsymbol{x}_{1,t}, \cdots, \boldsymbol{x}_{N,t}$:
$$\boldsymbol{x}_t{}^\top = \left(\boldsymbol{x}_{1,t}{}^\top, \cdots, \boldsymbol{x}_{N,t}{}^\top\right) \tag{5}$$
where t is a frame index and $N \geq 2$. An acoustic feature vector often consists of multiple streams: Mel Frequency Cepstrum Coefficients (MFCCs) and power parameters, or, static features and dynamic information. In the latter case, for example, a feature vector is divided into static and dynamic vectors.
2) Build an acoustic model M_i for each stream $\boldsymbol{x}_{i,t}$. An acoustic model M for original vectors $\boldsymbol{x}_t$ as well as a model for any combination of streams may be also required in the following processes.
3) In an i-th stream, perform speech recognition to estimate recognition accuracy (denoted by a_i). Obtained recognition transcriptions are used in 5) to 7).
4) Compare the accuracies among all the streams, and sort the streams according to the accuracies. In the following explanation, $a^{(j)}$ denotes the j-th highest accuracy score, and $i^{(j)}$ shows a stream index of $a^{(j)}$. For example, $a^{(1)}$ indicates the highest accuracy, then the index $i^{(1)}$ corresponds to the stream that achieved the highest performance.
5) When adapting an $i^{(1)}$-th stream, use only adaptation data $\boldsymbol{x}_{k,t}$ where $k = i^{(1)}$, as well as an acoustic model M_k for $\boldsymbol{x}_{k,t}$. The adaptation for the stream is then done using the adaptation data and the model, to obtain an adapted model $M'_{i^{(1)}}$.
6) For adaptation of an $i^{(2)}$-th stream, the following adaptation data $\boldsymbol{y}_{k,t}$ should be used:
$$\boldsymbol{y}_{k,t}{}^\top = \left(\boldsymbol{x}_{k_1,t}{}^\top, \boldsymbol{x}_{k_2,t}{}^\top\right) \tag{6}$$
where $k_1 = i^{(1)}$ and $k_2 = i^{(2)}$. An acoustic model M_y for the stream $\boldsymbol{y}_{k,t}$ obtained in 2) is also chosen. Using the data and the model, the model adaptation for the stream $\boldsymbol{y}_{k,t}$ is performed. Afterwards, the adapted model parameters $M'_{i^{(2)}}$ for $\boldsymbol{x}_{i^{(2)},t}$ are obtained.
7) Similarly, adapting $i^{(m)}$-th stream requires feature vectors $\boldsymbol{x}_{k_1,t}$, $\boldsymbol{x}_{k_2,t}$, $\cdots$, $\boldsymbol{x}_{k_m,t}$, as well as a corresponding acoustic model. The adapted model parameters $M'_{i^{(m)}}$ for the stream $\boldsymbol{x}_{i^{(m)},t}$ are extracted after the adaptation.
8) Through the last processes 5), 6) and 7), adapted model parameters M'_i for each stream are obtained. All the model parameters are subsequently assembled to construct an adapted model M'.
9) Using the adapted model M' and the original features $\boldsymbol{x}_t$, recognition is finally performed to obtain a recognition result.

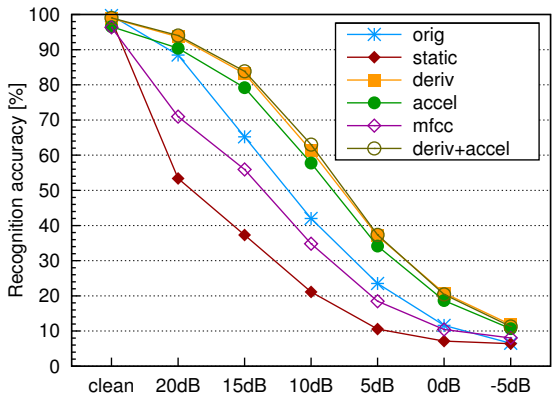

Fig. 2. Recognition accuracies of an original acoustic feature and its stream vectors (without adaptation).

TABLE I
ACOUSTIC STREAMS (VECTORS) USED IN SECTION III.

Name	Dim.	Parameter
orig	39	12 MFCCs, E, their Δ and $\Delta\Delta$
static	13	12 MFCCs and E
deriv	13	12 ΔMFCCs and ΔE
accel	13	12 $\Delta\Delta$MFCCs and $\Delta\Delta E$
mfcc	12	12 MFCCs
deriv+accel	26	12 ΔMFCCs, 12 $\Delta\Delta$MFCCs, ΔE and $\Delta\Delta E$

† E means a low-power coefficient.

III. EXPERIMENT

A. Database

In order to evaluate the proposed adaptation, Japanese connected-digit speech corpus CENSREC-1 (AURORA-2J) [6] was utilized. The training data set consists of 8,440 clean utterances made by 110 speakers (55 females and 55 males). The test data include not only clean but also noisy speech waveforms; eight kinds of noises (subway, babble, car, exhibition, restaurant, street, airport and station) were added to the clean speech at six SNR levels (20dB, 15dB, 10dB, 5dB, 0dB and −5dB), respectively. The test data set for each noise condition contains 1,001 utterances spoken by 104 speakers (52 females and 52 males). Each training or test utterance includes up to seven digits. Note that there are nine or ten utterances for each test speaker.

B. Experimental setup

Conventional HMM was employed as an acoustic model, which was built using the training data. An HMM was prepared for each word (digit or silence), having 16 states and 20 mixtures for digit, while 3 states and 36 mixtures for silence. The number of word HMMs was 13: "one", "two", ⋯, "nine", "zero", "oh", silence and short-pause. Any other experimental conditions were the same as those of the baseline system in CENSREC-1.

An original acoustic feature vector consisted of 12-dimensional MFCCs and a log power, their Δ and $\Delta\Delta$ (denoted by **orig**). This 39-dimensional feature vector was divided into three partial vectors: 13-dimensional static elements

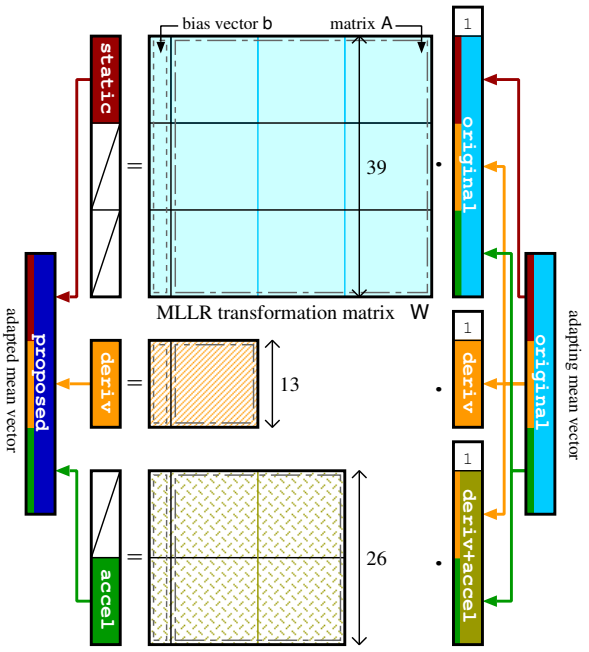

Fig. 3. Proposed MLLR adaptation matrix for 39-dimensional acoustic feature vectors (**static**, **deriv** and **accel**).

(**static**), 13 first-order dynamic coefficients (**deriv**) and 13 second-order dynamic parameters (**accel**), respectively. Table I summarizes the acoustic features used in this section.

C. Pre-recognition experiment

Figure 2 depicts recognition accuracies of every acoustic features (**orig**, **static**, **deriv** and **accel**) at six SNR levels as well as clean condition. For comparison, two additional features were also tested; **mfcc** had only 12 static MFCC coefficients, **deriv** and **accel** were combined into **deriv+accel**. The accuracy of **static** was much degraded in noisy environments mainly due to a static log-power coefficient. Suppressing the power (**mfcc**) recovered the accuracy to an extent, however, the accuracy was still lower than that of **orig**. On the other hand, dynamic features (**deriv** and **accel**) were relatively robust against noise. Furthermore, **deriv+accel** achieved the best performance among all the features (slightly better than **deriv**).

It is also found that the order of recognition accuracies was almost consistent in noisy environments:

$$\begin{aligned}\textbf{static} < \textbf{mfcc} < \textbf{orig} \\ < \textbf{accel} < \textbf{deriv} < \textbf{deriv+accel}\end{aligned} \quad (7)$$

In the following experiments, the fixed order denoted in (7) is adopted.

D. Model adaptation setup

The preliminary experiment shows that the most reliable stream is **deriv** followed by **accel**, whereas **static** is not powerful to distinguish digits in noisy environments. Therefore, our proposed adaptation technique employed the following strategy; mean vectors in a Gaussian distribution for

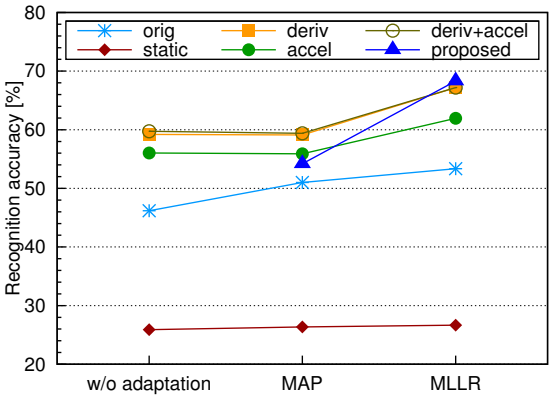

Fig. 4. Recognition accuracies with/without model adaptation.

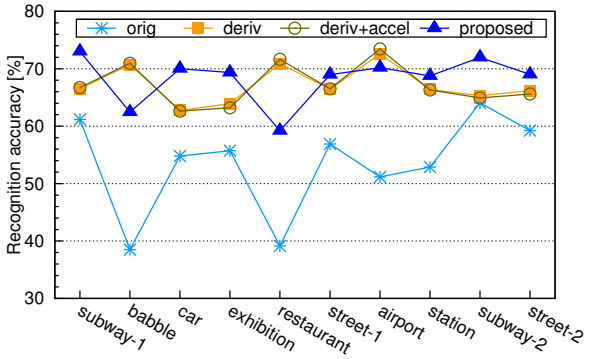

Fig. 5. Recognition accuracies of our proposed method and every streams after MLLR in noise conditions.

static were updated using 39-dimensional audio features (**orig**). Mean vectors for the first-order and second-order derivatives were adapted using **deriv** and **deriv+accel** parameters, respectively. Both unsupervised MAP-based and MLLR-based methods were investigated; all the utterances in each speaker were pre-recognized to obtain a recognition result, subsequently the adaptation was executed using the same speech data and the recognition result. Afterwards, the utterances were finally recognized using the adapted model. Regarding MLLR, global adaptation was applied in this paper. MLLR adaptation employed in the following experiment is illustrated in Figure 3.

E. Experimental result

Figure 4 shows adapted recognition results of each stream (**orig**, **static**, **deriv**, **accel** and **deriv+accel**) in addition to the proposed adaptation technique (denoted by **proposed**). Each entry means the average of recognition accuracies at the seven noise conditions. It is observed that **static** was obviously lower than the other streams. When conducting MAP, **orig** recovered the accuracy, however, there were few improvements in **deriv** and **accel**. This may be caused due to lack of adaptation data. In contrast, the performances of MLLR results were successful. For example, **deriv+accel** achieved 21.0% and 13.8% absolute improvements compared to **orig** without/with MLLR, respectively.

Regarding the proposed scheme, 3.4% relative error reduction on average was achieved in MLLR, compared with **deriv+accel**. Since this might seem not to be a significant improvement, we conducted a further analysis. Figure 5 shows MLLR-applied recognition performances of **proposed** as well as its source streams **orig**, **deriv** and **deriv+accel** in ten noise conditions. **proposed** succeeded in seven conditions (maximally 19% relative error reduction from **deriv** or **deriv+accel**), however, it could not recover the accuracy in the other cases (babble, restaurant and airport). In such the cases, it is found that **orig** performance is quite low; the accuracy was sometimes below zero at low SNR circumstances. This means the adaptation did not work well due to corrupted transcriptions. Since the proposed method used the degraded model for static coefficients, the performance fell down from the **deriv** and **deriv+accel**. It is thus crucial to use more reliable transcription (e.g. **deriv+accel**) to adapt a corrupted stream (e.g. **static**). In addition, it seems that the proposed scheme worked well in relatively stationary noises (e.g. subway and car), on the other hand, the method suffered from non-stationary environments (e.g. babble and restaurant). In such the environments, incremental or frame-by-frame adaptation may be effective.

IV. Conclusion

This paper proposes a multi-stream model adaptation; an input feature vector is divided into several streams. When adapting, model parameters in each stream are updated using streams where their performances are superior to the current stream. Recognition experiments for static and dynamic streams were conducted, followed by adaptation experiments. The recognition accuracy drastically increased from 46.2% (baseline) to 68.3% (using proposed adaptation). From the results, it is concluded that the superior audio stream is helpful to adapt the inferior audio stream in model adaptation, as same as the multi-modal model adaptation we previously reported.

As our future work, we would like to deeply explore the mechanism and factor of further improvement. For example, instead of choosing appropriate streams, introducing stream weight factors in model adaptation might be investigated. Estimation of stream confidence in order to determine the stream order or to choose effective streams is also expected.

References

[1] J.L.Gauvain et al., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Trans. on Speech and Audio Processing, vol.2, no.2, pp.291-298 (1994).
[2] C.J. Leggetter et al., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, pp.171-185 (1995).
[3] R.Kuhn, et al., "Eigenvoices for speaker adaptation," Proc. ICSLP'98, pp.1771-1774 (1998).
[4] C.Yang et al., "Static and dynamic spectral features: their noise robustness and optimal weights for ASR," Proc. ICASSP2005, pp.241-244 (2005).
[5] S.Tamura et al., "Audio-visual interaction in model adaptation for multi-modal speech recognition," Proc. APSIPA ASC 2011, Thu-PM.PS2.7 (2011).
[6] S.Nakamura et al., "AURORA-2J: an evaluation framework for Japanese noisy speech recognition," IEICE Trans. on Information and System, vol.E88-D, no.3, pp.535-544 (2005).