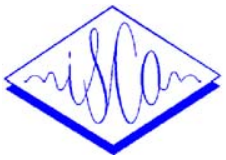
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Robust ASR Model Adaptation by Feature-Based Statistical Data Mapping

Xuechuan Wang *Douglas O'Shaughnessy*

INRS-EMT, University of Quebec
800 de la Gauchetiere West
Montreal, Quebec, H5A 1K6, Canada
wwang, dougo@inrs-emt.uquebec.ca

10.21437/Interspeech.2004-12

Abstract

Automatic speech recognition (ASR) model adaptation is important to many real-life ASR applications due to the variability of speech. The differences of speaker, bandwidth, context, channel and et al. between speech databases of initial ASR models and application data can be major obstacles to the effectiveness of ASR models. ASR models, therefore, need to be adapted to the application environments. Maximum Likelihood Linear Regression (MLLR) is a popular model-based method mainly used for speaker adaptation. This paper proposes a feature-based statistical Data Mapping (SDM) approach, which is more flexible than MLLR in various applications, such as different bandwidth and context. Experimental results on the TIMIT database show that ASR models adapted by the SDM approach have improved accuracy.

1. Introduction

To handle the diverse nature of variation found in speech, it is important to adapt ASR models to changes to maintain ASR accuracy throughout. It has been feasible recently to establish initial speaker-independent (SI) ASR models for context-dependent (CD) phonemic models, using the standard hidden Markov model (HMM) approach, based on analysis of much recorded speech. For increased accuracy and real-life ASR applications, the initial models must be adapted, as their precision is insufficiently focused for specific applications. Model adaptation can be achieved in two ways: one is model-based direct adaptation, i.e., adapt ASR models directly with the environmental compensation data. The other is feature-based indirect adaptation, which maps the adaptive data into the SI environment so that the mapped adaptive data have minimum difference with the SI model environment and maximum information of the adaptive environment. The ASR models are then retrained with the mapped adaptive data.

MLLR is a commonly used model-based adaptation technique, which assumes that the parameters of the Gaussians (mean and variance) in ASR models are transformed by an affine transformation into parameters that better match the adaptation data [1]. MLLR is suitable to compensate the time-invariant changes, such as speaker, channel or additive noise. Practical constraints, however, usually severely limit the effectiveness of MLLR adaptation. For example, the lack of enough adaptation data, different system configurations and speech context may lead to the failure of MLLR adaptation. This paper proposes a feature-based non-linear SDM approach. The SDM approach assumes that speech observations are generated by subsets of mutually related random sources. The relationship/model between random sources can be established by maximum likelihood criterion. Thus speech observations from an adaptive environment can be mapped through the relationship/model into any desired environment without heavy loss of the original information.

The rest of this paper is organized as follows: Section 2 introduces the framework of the SDM approach and varations of the SDM approach. In Section 3 we give a description of the database and experimental setup. The results of the experiments are given in the same section. In Section 4 we conclude with the results obtained.

2. Feature-Based Statistical Data Mapping

2.1. General Principle

Consider Figure 1, and suppose that speech signals are generated by a set of random sources $\mathcal{S} = \{s_1, s_2, \cdots, s_L\}$. Each random source has a probability density function, characterized by a mean and a covariance matrix [2]. Speech signals are such random signals that are emitted from certain sources with the highest probability. Both original and adaptive speech signals can be considered as random signals emitted from subsets of $\mathcal{S}$, Θ and Λ, respectively. The adaptive sources Λ are related to the original sources Θ by conditional probability $P(\Theta|\Lambda)$, as shown in Figure 1, which indicates the contributions of narrow-band sources to wide-band speech. It is thus feasible to estimate the wide-band speech through

This work is supported by the Canadian government under the NSERC Strategic Partnerships Program.

a statistical recovery function (SRF) [2, 3].

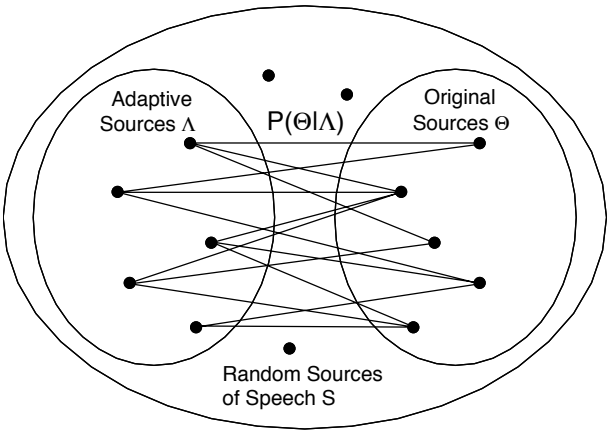


Figure 1: Adaptive sources are related to original sources by conditional probability $P(\Theta|\Lambda)$.

2.2. The Statistical Recovery Function

Assume that a set of adaptation speech observations $\mathcal{X} = \{x_1, x_2, \cdots, x_T\}$ is generated by N adaptive random sources $\Lambda = \{\lambda_i\}, 1 \leq i \leq N$ and the corresponding original observation set $\mathcal{Y} = \{y_1, y_2, \cdots, y_T\}$ by M random sources $\Theta = \{\theta_j\}, 1 \leq j \leq M$. Given their underlying sources, the conditional probability density functions of the original and adaptive observations at time t, y_t and x_t, are $p(y_t|\theta_j)$ and $p(x_t|\lambda_i)$, respectively. The cross-correlation probability of original source θ_j contributing to y_t, given the adaptive source λ_i contributing to x_t, is defined as $\alpha_{ij} = p(\theta_j|\lambda_i)$. Let the set of cross-correlation probabilities be $A = \{\alpha_{ij}\}$. The SRF of $\mathcal{Y}$ from $\mathcal{X}$ is defined as:

$$\mathcal{Y} = f(\mathcal{X}, A, \Lambda, \Theta). \tag{1}$$

Given a pair of observation sequences $(\mathcal{X}, \mathcal{Y})$, where

$$(\mathcal{X}, \mathcal{Y}) = \{(x_t, y_t)\}, 1 \leq t \leq T, \tag{2}$$

the parameter set $\{A, \Lambda, \Theta\}$ can be optimized by maximizing the likelihood, $p(\mathcal{X}, \mathcal{Y})$, through the EM algorithm [4]. $p(\mathcal{X}, \mathcal{Y})$ is equivalent to $p(\mathcal{X}, \mathcal{Y}, A, \Lambda, \Theta)$ in this case and can be replaced by it:

$$\begin{aligned} p(\mathcal{X}, \mathcal{Y}) \quad &\Leftrightarrow p(\mathcal{X}, \mathcal{Y}, A, \Lambda, \Theta) \\ &= \textstyle\prod_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} p(x_t, y_t, \lambda_i, \theta_j), \end{aligned} \tag{3}$$

where $p(x_t, y_t, \lambda_i, \theta_j)$ is the joint pdf for the speech pair (x_t, y_t) and individual sources λ_i and θ_j at time t. Considering that y_t is independent of x_t and λ_i and x_t independent of θ_j, $p(x_t, y_t, \lambda_i, \theta_j)$ can be calculated as follows:

$$p(x_t, y_t, \lambda_i, \theta_j) = p(y_t|\theta_j)\alpha_{ij}p(x_t|\lambda_i)p(\lambda_i). \tag{4}$$

In the E-step, the auxiliary function $Q(\Phi|\Phi_0)$ is defined as:

$$Q(\Phi|\Phi_0) = E(\log p(\Phi, \mathcal{X}, \mathcal{Y}, s_k)|\Phi_0). \tag{5}$$

$Q(\Phi|\Phi_0)$ is calculated over all the state paths s_k, where

$$s_k = \{[\lambda_i^k, \theta_j^k]|_{t=1}, [\lambda_i^k, \theta_j^k]|_{t=2}, \cdots, [\lambda_i^k, \theta_j^k]|_{t=T}\} \tag{6}$$

is the k-th state path, $i = 1, \cdots, N$, $j = 1, \cdots, M$, $k = 1, 2, \cdots, (NM)^T$, $\Phi = \{A, \Lambda, \Theta\}$ and Φ_0 is the initial parameter set. The conditional joint pdf of λ_i and θ_j at time t is obtained by

$$p(\lambda_i, \theta_j, t|\mathcal{X}, \mathcal{Y}) = \frac{p(\lambda_i, \theta_j, x_t, y_t)}{\sum_{n=1}^{N} \sum_{m=1}^{M} p(\lambda_n, \theta_m, x_t, y_t)}. \tag{7}$$

The conditional pdfs of λ_i and θ_j are:

$$p(\lambda_i, t|\mathcal{X}, \mathcal{Y}) = \frac{\sum_{j=1}^{M} p(\lambda_i, \theta_j, x_t, y_t)}{\sum_{n=1}^{N} \sum_{m=1}^{M} p(\lambda_n, \theta_m, x_t, y_t)} \tag{8}$$

$$p(\theta_j, t|\mathcal{X}, \mathcal{Y}) = \frac{\sum_{i=1}^{N} p(\lambda_i, \theta_j, x_t, y_t)}{\sum_{n=1}^{N} \sum_{m=1}^{M} p(\lambda_n, \theta_m, x_t, y_t)}. \tag{9}$$

In the M-step, the auxiliary function $Q(\Phi|\Phi_0)$ is maximized subject to the constraint $\sum_{j=1}^{M} \alpha_{ij} = 1$. Given that both the adaptive and original sources are Gaussian sources with diagonal covariance matrices, the source parameters are obtained by:

$$\alpha_{ij} = \frac{\sum_{t=1}^{T} p(\lambda_i, \theta_j, t|\mathcal{X}, \mathcal{Y})}{\sum_{t=1}^{T} p(\lambda_i, t|\mathcal{X}, \mathcal{Y})}, \tag{10}$$

$$p(\lambda_i) = \frac{1}{T} \sum_{t=1}^{T} p(\lambda_i, t|\mathcal{X}, \mathcal{Y}), \tag{11}$$

$$\mu_{\lambda_i} = \frac{\sum_{t=1}^{T} x_t p(\lambda_i, t|\mathcal{X}, \mathcal{Y})}{\sum_{t=1}^{T} p(\lambda_i, t|\mathcal{X}, \mathcal{Y})}, \tag{12}$$

$$\Sigma_{\lambda_i} = \frac{\sum_{t=1}^{T} (x_t - \mu_{\lambda_i})^2 p(\lambda_i, t|\mathcal{X}, \mathcal{Y})}{\sum_{t=1}^{T} p(\lambda_i, t|\mathcal{X}, \mathcal{Y})}, \tag{13}$$

$$\mu_{\theta_j} = \frac{\sum_{t=1}^{T} y_t p(\theta_j, t|\mathcal{X}, \mathcal{Y})}{\sum_{t=1}^{T} p(\theta_j, t|\mathcal{X}, \mathcal{Y})}, \tag{14}$$

$$\Sigma_{\theta_j} = \frac{\sum_{t=1}^{T} (y_t - \mu_{\theta_j})^2 p(\theta_j, t|\mathcal{X}, \mathcal{Y})}{\sum_{t=1}^{T} p(\theta_j, t|\mathcal{X}, \mathcal{Y})}, \tag{15}$$

where $p(\lambda_i)$ is the a priori probability of source λ_i, μ_{λ_i} and Σ_{λ_i} are the mean and the covariance matrix of source λ_i, and μ_{θ_j} and Σ_{θ_j}, for source θ_j. The optimized parameter set Φ can be obtained through a standard Levinson-Durbin recursive algorithm [5].

Since the absolute energy of the input speech signals varies in different cases, it has little direct utility in calculating the statistical parameters of sources. The influence

of energy level can be effectively removed by using an energy-normalized input. The absolute energy value of original speech can be recovered, correspondingly, from the adaptive speech energy and the energy ratio. The energy ratio, ε, is defined as:

$$\varepsilon = \left(\frac{\mathcal{E}(y_t)}{\mathcal{E}(x_t)} \right)^{1/2}. \quad (16)$$

The energy ratio under a given original source θ_j, ε_{θ_j}, is obtained by:

$$\varepsilon_{\theta_j} = \frac{\sum_{t=1}^{T} \left(\frac{\mathcal{E}(y_t)}{\mathcal{E}(x_t)} \right)^{1/2} p(\theta_j | x_t, y_t)}{\sum_{t=1}^{T} p(\theta_j | x_t, y_t)}. \quad (17)$$

2.3. Adaptive Data Mapping

The SRF defined in Equation (1) can be realized by various criteria to map adaptive speech to original speech $\mathcal{Y}$. Minimum mean square estimation (MMSE) is a commonly used criterion. An estimation directly derived from a MMSE criterion is the conditional expection, which is defined as:

$$\hat{Y} = E(\mathcal{Y}|\mathcal{X}) = \int_T \mathcal{Y} p(\mathcal{Y}|\mathcal{X}) d\mathcal{Y} = \int_T \mathcal{Y} \prod_{t=1}^{T} p(y_t|x_t) d\mathcal{Y}. \quad (18)$$

Without loss of generality, assume that the observation pairs (x_t, y_t), $t = 1, \cdots, T$ are independent of each other. Thus,

$$\hat{Y} = \{ \int y_t p(y_t|x_t) dy_t \}, \quad t = 1, \cdots, T. \quad (19)$$

Given N adaptive sources and M original sources, the integral in Equation (19) is obtained by:

$$\int y_t p(y_t|x_t) dy_t = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} \mu_j \alpha_{ij} p(x_t|\lambda_i) p(\lambda_i)}{\sum_{i=1}^{N} p(x_t|\lambda_i) p(\lambda_i)}. \quad (20)$$

The original speech obtained from Equations (19) and (20) are energy-normalized speech vectors, which are not ready for use before the original energy is recovered. The original energy can be obtained from the energy ratio and adaptive energy as follows:

$$(\mathcal{E}_{y_t})^{1/2} = (\mathcal{E}_{x_t})^{1/2} \sum_{j=1}^{M} \varepsilon_{\theta_j} p(\theta_j, t | \mathcal{X}, \mathcal{Y}). \quad (21)$$

2.4. Variations of SDM

SDM can be easily varied depending on the clustering of the random sources. In this paper, we propose two variations, one is gender-specific SDM and the other is phonemic-specific SDM.

2.4.1. Gender-Specific (GS) SDM

Speech signals from male and female speakers have different frequency structure. Gender specification will thus assign the SDM model better discriminative power over fine structure of speech features. In GS SDM variation, two sets of random sources are set up by gender and trained separately.

2.4.2. Phonemic-Specific (PS) SDM

It is clear that the SDM model will gain a better match to the speech features if it is built up by phonemes. Thus in PS SDM, random speech sources are clustered by phonemes. The PS SDM model is then trained on each phoneme.

3. Experiments and Results

3.1. Database and ASR System

The model adaptation experiments are conducted on the TIMIT [6] and NTIMIT [7] databases. The TIMIT corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. NTIMIT was collected by transmitting all 6300 original TIMIT utterances though various channels in the NYNEX telephone network and redigitizing them. In the experiments, TIMIT data is used as the original database for SI models. NTIMIT data is used as adaptive data to provide a different speaker, bandwidth and context environment.

The HTK-based speech recognition system [8] is used throughout the experiments. HTK is a Hidden Markov Model (HMM)-based speech recognition system and designed for both isolated and continuous speech recognition. A continuous whole-word-based speech recognition system is built in our experiments. The system uses 5-state HMMs for each of 47 monophones. Each state of the HMMs has a 15-mixture Gaussian model (GMM).

The speech features used in the ASR system are 39 MFCC features, which include 13 MFCCs, 13 Deltas and 13 Accelerations.

3.2. Experiment Configuration

Three experiments are conducted in our test on SDM. They are:

- **Speaker Adaptation** – the baseline SI models are trained on dr1 and dr3-dr8 of the training set of the TIMIT database. The SDM model are obtained from dr2 of the training set of both the TIMIT and NTIMIT databases. The SI model and adapted model are tested on the dr2 of the test set of the TIMIT database.

- **Bandwidth Adaptation** – the baseline SI models are trained on the training set of the TIMIT database. The SDM models are obtained from the training set of both the TIMIT and NTIMIT databases. The ASR model and adapted model are tested on the test set of the TIMIT database.
- **Speech Context Adaptation** – utterances of SA1 and SA2 are extracted as the new speech context. All the SA1 and SA2 utterances are taken out from the training set to obtain an ASR model without the context of SA1 and SA2. The SA1 and SA2 utterances in the NTIMIT database are used to provided an adptive SA1 and SA2 context. The test set is composed by the SA1 and SA2 utterances from the test set of the TIMIT database.

3.3. Experimental Results

The number of states (N and M in Equation 3) has a direct influence on the performance of the SDMs. Table 1 shows the results of GS SDM and PS SDM with different number of states on bandwidth adaptation test. The results show that increasing state number can improve the performance of the SDMs. However, the performance of the SDMs may be degraded if the number is too high.

States	64	128	3/PH	5/PH	8/PH	11/PH
GS SDM	29.66	28.97	–	–	–	–
PS SDM	–	–	35.91	31.17	28.05	30.20

Table 1: *WER results of GS/PS SDMs with different number of states on bandwidth adaptation test (N=M).*

Table 2 shows the word error rate (WER) results of the three experiments. In speaker and bandwidth adaptation tests, general SDM does not give improvement to the WER over a baseline SI model. GS SDM gives a small improvement in speaker adaptation tests, but not in bandwidth adaptation tests. Both PS SDM and combined PS and GS SDM give improvement over SI models for the two tests and have better performances than MLLR in speaker adaptation test. In speech context adaptation tests, all the SDM models have significant improvements over a baseline SI model. PS SDM and combined PS and GS SDM have especially high improvements.

Model	States	Speaker	Bandwidth	Context
Baseline	–	27.14	28.89	42.80
MLLR	–	26.83	–	–
General SDM	128	27.88	29.46	11.37
GS SDM	128	26.78	28.97	11.14
PS SDM	8/PH	25.41	28.05	10.49
PS & GS SDM	8/PH	25.37	26.95	10.49

Table 2: *Adaptation WER results (N=M).*

4. Conclusions

In this paper, we propose a feature-based SDM approach and its variations for robust ASR model adaptation. The results show that variations such as PS SDM and combined PS and GS SDM have better performances than general SDM. Compared to MLLR, which is a model-based adaptation technique, SDM has better performance in speaker adaptation. It is also flexible to deal with non-environmental changes such as bandwith and context changes. The SDM approach has the potential of developing into accurate variations in specified application environments. This property is highly important to real-life ASR applications nowadays.

5. References

[1] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M. Picheny, "Speaker clustering and transformation for speaker adaptation in speech recognition sytems", *IEEE Trans. Speech and Audio Processing*, Vol. 6, pp 71-77, Jan 1998.

[2] Y.M. Cheng, D. O'Shaughnessy and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, 1994, pp. 544-548.

[3] K.Y. Park, and H.S. Kim, "Narrowband to wideband conversion of speech using GMM-based transformation", *ICASSP'00*, Istanbul, June 2000, pp. 1843-1846.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Annual Royal Statistical Society*, 1977, pp. 1-38.

[5] S.M. Kay, *Modern Spectral Estimation: Theory and Application*, Englewood Cliffs, NJ: Prentice Hall, 1988.

[6] W. Fisher, G. Doddington and K. Goudie-Mashall, The DARPA speech recognition research database: specifications and status, *Proceedings DARPA of Speech Recognition Workshop*, 1986, pp. 93-99.

[7] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database", *Proceedings of ICASSP'90*, April 1990, pp. 109-112

[8] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book*, Cambridge University Speech Group, July 2000.