Modeling Sound Changes in Mandarin Spontaneous Speech Using Deleted Interpolation of Mixture Component Weights *

LIU Yi[1]， HE Lei[2], ZHENG Fang[1]
1. Center for Speech and Language Technology, Division of Technical Innovation and Development
Tsinghua National Laboratory for Information Science and Technology, Beijing
2. Toshiba (China) Research and Development Center, Beijing

文　摘：The high error rate of recognition accuracy in spontaneous speech is due in part to the poor modeling of pronunciations variations. An analysis of the acoustic data reveals that the variations include both phone changes and sound changes. Sound changes are the variations within the phoneme, such as nasalization, centralization, voiceless, voiced, etc. Sound changes are flexible and include diacritics that have to be explicitly hand-labeled by linguists. Annotating such corpus is time consuming and the available hand-labeled samples of sound changes are very limited. In this paper, based on standard phonetic unit inventory, we use dynamic programming alignment together with data-driven method to extend the phone set automatically for sound change description. We propose using deleted interpolation to interpolate baseline models and the more refined, but less well-trained sound change models, with the goal of improving the robust ability of sound change models to cover the diversity of sound variations in spontaneous speech. The effectiveness of this approach is evaluated on the 1997 Hub4NE Mandarin Broadcast News Corpus (1997 MBN) with different styles of speech. It gives a significant 1.98% absolute syllable error rate reduction in spontaneous speech. Additional 1.04% absolute syllable error rate reduction is achieved compared to Gaussian mixture sharing method.

关键词：Sound Changes; Deleted Interpolation; Hand-labeled
中图分类号：TN912.34

1. Introduction

The maximum likelihood (ML) criterion is widely used in speech recognition for HMM parameters training [1]. If a large amount of training data is provided, good estimations of these parameters can be obtained. However, when only a limited amount of data is available, parameters of some HMM units would be poorly estimated which leads to recognition performance degraded. In Mandarin spontaneous speech, there is a large amount of pronunciation variations because of the casual way of speaking. The variations can be classified into phone changes and sound changes [2]. Phone changes, are the replacement of a phoneme by another alternate phone, such as ‘b’ being pronounced as ‘p’. Sound changes, are the variations within the phoneme such as nasalization, centralization, voiceless and voiced. Both complete changes and partial changes are very common in spontaneous Mandarin speech. In [3], phone level transcriptions were studied for labeling a spontaneous Mandarin speech corpus – CASS. SAMPA_C labels were generated to annotate sound variability. SAMPA-C is a phone level transcription, which includes standard symbols for consonants and vowels, initials and finals, sound variability and spontaneous phenomena. In [4], pre-defined Generalized Initial/Final (GIF) units derived from SAMPA-C unit set were used to represent pronunciation variations in Mandarin speech. The GIF set includes the standard Initial/Final set as well as the extended phone set. Using GIF units is to augment the phone inventory artificially with the goal of labeling the alternative pronunciations. However, we discovered that in most cases,

*基金项目：The work is partially supported by Toshiba and Tsinghua University Joint Project.
作者简介：LIU Yi (1972), Male (Han), Associate Professor.
通讯联系人：LIU Yi, Associate Professor, eeyliu@tsinghua.edu.cn

there are only a limited number of training samples for the non-canonical part of the GIF set, thus the parameters of GIF models cannot be robustly estimated. Meanwhile, it is very time consuming to label spontaneous speech by hand. Using hand-defined symbols based on GIF is inadequate for identifying a lot of variations that cannot be distinguished consistently by phoneticians. In other words, phoneticians have low confidence with the extended phonetic units, especially when sound changes are commonly existed in spontaneous speech. In addition, the amount of available training samples of extended phonetic units is always insufficient for robust acoustic model training. On the other hand, generating acoustic sub-word unit (ASU) that is learned solely from training data is able to model part of sound changes, however, the use of ASU always makes it difficult to define a pronunciation dictionary [5]. Another approach of merely using triphone units can only model phonetic confusions caused by contextual effects, but cannot model the inherent phonetic confusions arising from the accent or other speech effects (e.g., the confusion between 'z' and 'zh', 'n' and 'l' in Mandarin speech).

Deleted interpolation (DI) is regarded as one of the most powerful smoothing techniques to improve the performance of acoustic models as well as language models in speech recognition [6, 7]. It is often necessary to combine well-trained general models with less well-trained but more refined models. Deleted interpolation has been successfully used for this purpose for both discrete and semi-continuous HMM [6]. In conventional approach, DI is used for smoothing HMM parameters such as mean, variance between general and refined models [6, 7]. However, for continuous HMMs, such as those used in our experiments, it is more convenient to use DI on mixture weights other than mean and variance, as the latter are harder to estimate.

In this paper, we propose an approach of interpolating well-trained canonical models with those less well-trained refined models in order to strengthen the robustness of the acoustic model to cover sound changes. The deleted interpolation method is used to smooth the mixture component weights rather than all HMM parameters. Meanwhile, instead of manual derivation of phonetic unit inventory and data-driven generation of ASU units, the phonetic units for refined models are based on the standard phonetic unit inventory and are learned from the samples obtained through DP alignment between the canonical and alternate phone strings. An iterative estimation-maximization (EM) procedure is used for the interpolation weight estimation. In order to improve the efficiency of DI, we generate deleted models by sharing Gaussian mixture components between well-trained general model and less well-trained refined model. The weights of Gaussian mixtures in deleted HMM are governed by interpolation coefficient.

This paper is organized as follows. Section 2 introduces sound changes in Mandarin spontaneous speech. Section 3 describes the method of automatic phone set generation. In Section 4, we explain the deleted interpolation mechanism and how to share mixture components between general and refined models. The experimental results are given in Section 5, and we conclude in Section 6.

2. Sound Changes

Pronunciation in spontaneous speech is very flexible. There are a lot more phonetic shifts, reduction and assimilation, duration changes, tone shifts, etc. compared to read speech and planned speech. Linguistic knowledge and empirical results show that pronunciation variations in Mandarin can be classified into two types: phone changes and sound changes. Phone changes are the replacement of a canonical phone by another canonical phone, such as *ai* being pronounced as *ei*. Sound changes are variations within the same phonemes, such as nasalization, centralization, voiceless, voiced, rounding, syllabic, pharyngealization, and aspiration. For example, *f* can change into *f_v* (voiced), *ts* to *ts_v* or *ts_h*, these changes are called diacritics. Phone changes can be modeled by canonical phone models and trained in the normal way. Sound changes are a little more difficult to model as this diacritic set of phones can only be trained by the samples labeled initially by humans. In other words, sound changes are occurred within the phoneme and are caused by sound variability [3]. They are very flexible and a lot less clear-cut than previously assumed and cannot be modeled by mere representation in alternate or concatenation of phone units [2, 8]. When partial changes occur, a phone is not completely substituted, deleted or inserted. Table 1 illustrates the examples of description of canonical pronunciations labeled with Initial/Final units and the relevant sound changes labeled with SAMPA_C.

Table 1. Examples for sound changes description

Pinyin	SAMPA_C	Comments
Z	/ts/	Canonical
Z	/ts_v/	Voiced
Z	/ts`/	Changed to '*zh*'
z	/ts`_v/	Changed to voiced '*zh*'
e	/7/	Canonical
e	/7`/	Retroflexed, or changed to '*er*'
e	/@/	Changed to /@/ (a GIF)

Sound changes are variations within the phoneme. When sound changes occur, a phone is not completely substituted, deleted or inserted. Therefore, the transcriber agreement on spontaneous speech is much lower than that on read speech. The different transcriber agreement rates suggest that when sound changes occur, the transcribers who are forced to use a categorical label from the limited phonetic inventory may end up choosing different labels at the phone level representation. A similar situation can be found in ASR tasks. Due to the effect of sound changes the degree of phonetic confusions is increased in spontaneous speech, which leads to low discriminative power of acoustic models, resulting in the degradation of recognition performance.

3. Automatic Phone Set Extension

The phone set can be extended to describe variants by either using hand-defined symbols based on phonological knowledge or using data-driven method [2, 5]. However, it has been shown that increasing phone set by using hand-defined symbols is very time consuming and insufficient in identifying a lot of phonetic confusions which cannot be distinguished consistently by phoneticians. We use the standard phonetic unit inventory together with data-driven method to extend phone set automatically. The procedure of generating the initial set of extended phonetic units is illustrated in Fig. 1 and is as follows:

(1) *Generate baseform transcriptions*. The baseform transcriptions can be obtained by looking up a canonical word-to-phoneme dictionary. When a word has multiple pronunciations, the correct phoneme concatenation is selected by transcribers.

(2) *Generate surface form transcriptions*. In general, hand-labeled transcriptions are used as the surface form representation. However, due to a limited amount of available hand-labeled transcriptions, they are always insufficient for acoustic model training. Alternatively, we use automatic phone recognition method described in [9] to get the most likely phonetic sequence which is regarded as the surface form transcriptions.

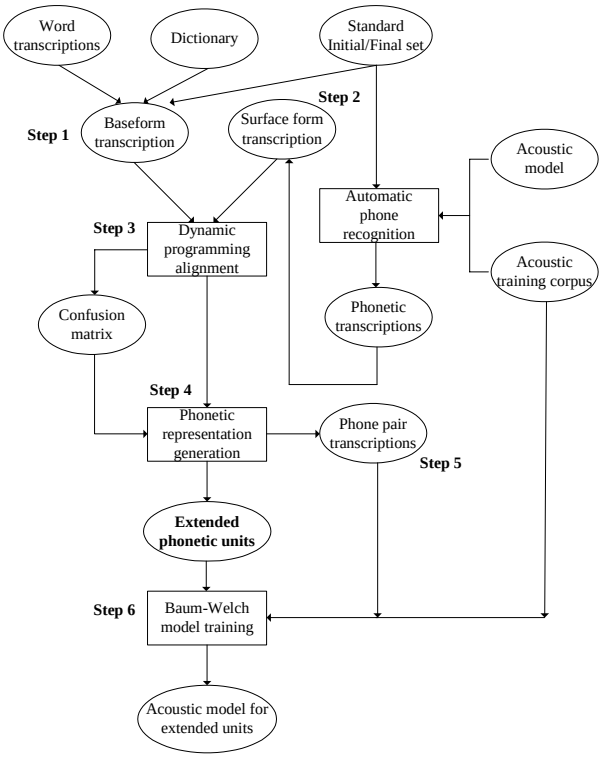

Figure 1: Flowchart for automatic phone set extension.

(3) *Align baseform and surface form transcriptions.* A DP alignment tool with flexible local edit distance measure [9] is used for baseform and surface form alignment.

(4) *Obtain the inventory of extended phone units (EPU).* The generation of initial inventory of extended phone units is based on the mapped baseform/surface form phone pairs. Through DP alignment, if the subword unit in the baseform string maps to another different subword unit in the surface form string, we combine these two units to form a phone pair as an extended phone unit, such as 'b_d'. To circumvent the sparse problem, we use log likelihood ratio test a confidence measure to discard the extended phone units [10].

(5) *Generate phone-level transcriptions in terms of standard phone units and extended phone units.*

If a phoneme in the baseform transcription has a relevant alternate phone in the surface form transcription, and this phoneme-to-phone pair can be found in the inventory of extended phone units, the phoneme in the baseform string is replaced by the extended phone unit.

(6) *Generate refined acoustic model for extended phone units.* The initial acoustic parameters of refined models with extended phone set description are cloned from their corresponding baseform models (e.g., the initial acoustic parameters of 'b_d' and 'b_p' are from those of 'b') and re-estimated using the Baum-Welch (BW) algorithm with the transcriptions generated from Step 5.

4. Deleted Interpolation and Sharing Mixture components

Due to the limitation of hand-labeled data with sound change labeling on spontaneous corpus, the acoustic model of sound change units cannot be trained robustly with limited number of training samples. We propose using deleted interpolation together with state-level Gaussian mixture sharing at each iteration step to improve the robustness of sound-change models. The models are regarded as less well-trained but more refined models, whereas baseline models with standard phone se units are less refined but well trained models.

4.1 Deleted Interpolation

Deleted Interpolation is used to smooth less well-trained but more refined models with well-trained general model. With cross-validation data, an iterative estimation-maximization (EM) procedure is used for estimating the interpolating weights. In discrete HMMs, output distributions can be interpolated directly [6], while we focus on continuous density HMMs in which each observation probability distribution is represented by a mixture Gaussian density. Define $P_i^{Detail}(\cdot)$ as distributions of refined but less well-trained model and $P_i^{General}(\cdot)$ is general well-trained model. A general deleted interpolation equation is as follows:

$$P_i^{DI}(\cdot) = \lambda_i P_i^{Detail}(\cdot) + (1-\lambda_i) P_i^{General}(\cdot) \qquad (1)$$

where $P_i^{DI}(\cdot)$ is the mixture components after deleted interpolation and λ_i is the interpolation weight. Based on the discussion in Section 3, $P_i^{General}(\cdot)$ refers to baseline acoustic model, so it can be rewritten as $P_i^{IF}(\cdot)$. $P_i^{Detail}(\cdot)$ refers to extended phone unit (EPU) model, that is $P_i^{EPU}(\cdot)$

The purpose of deleted interpolation is to obtain the interpolating weight λ_i for each EPU model. λ_i is estimated by cross-validation. For each needed to be interpolated unit, we assigned different interpolation weight λ_i. The algorithm for estimating λ_i is as follows:

1. Given a set of training data denoted as B, divide B into M sets, $B_1, B_2, \cdots, B_M$,
2. Train $P_i^{IF}(\cdot)$ and $P_i^{EPU}(\cdot)$ model from each combination of $M-1$ parts with EM algorithm, the residual part j is reserved as the deleted part for cross-validation.
3. Define the observation in deleted set as $O = \{o_1, o_2, \cdots, o_K\}$. Perform phone recognition on each deleted part j using the $P_i^{EPU-j}(\cdot)$ models, save time sequence information during phone recognition.
4. Align phone recognition sequence (observed) with baseform sequence (canonical) using the DP alignment. Find the canonical unit that has been aligned with alternative unit i. According to the time sequence information in Step.3, record its corresponding phone observation in the deleted set.
5. Let $c(o_k, P_i^{EPU-j}(\cdot) | B_m)$ be the counts of occurrences of observation o_k and model $P_i^{EPU-j}(\cdot)$ in deleted set B_m.
6. Applying an iterative EM method for estimating

the interpolation weights. Define the count for λ_i as η_i where

$$\eta_i = \sum_{j=1}^{M}\sum_{k=1}^{K}\{c(o_k, P_i^{EPU-j}(\cdot)\mid B_j) \times \frac{\lambda_i \cdot P_i^{EPU-j}(o_k)}{\lambda_i \cdot P_i^{EPU-j}(o_k)+(1-\lambda_i)P_i^{IF-j}(o_k)}\} \quad (2)$$

Update the interpolation weights:

$$\lambda_i^{new} = \frac{\eta_i}{\sum_{j=1}^{M} c(\cdot)} \quad (3)$$

According to Eq.(1), share Gaussian mixture components between EPU and IF models to generate a new interpolated model. At each iteration, calculate the log likelihood LE_i with deleted set:

$$LE_i = \sum_{j=1}^{M}\sum_{k=1}^{K} c(o_k, P_i^{EPU-j}(\cdot)\mid B_j)\log(P_i^{DIEPU-j}(o_k))$$

If LE_i converges, which means the interpolation weight is also converged, we stop iteration. If not, substitute $P_i^{EPU-j}(\cdot)$ by $P_i^{DIEPU}(\cdot)$, go to Step 6.

4.2 Gaussian Mixture Components Sharing

From Eq.(1), we can see that the Gaussian mixture components of deleted model consist of mixture components of original detailed and general models. The mixture weights of deleted model are governed by interpolation weights. After sharing mixture components, the deleted models must be smoothed by original well-trained models. In addition, other HMM parameters of deleted models can be updated at each iteration during EM training procedure. Let

$$P_i^{EPU}(\cdot) = \sum_a w_{ia}\mathrm{N}(o;\mu_i,\sum_i)$$

$$P_i^{IF}(\cdot) = \sum_b w_{ib}\mathrm{N}(o;\mu_i,\sum_i)$$

a, b are the total mixture numbers of one state, and w_{ia}, w_{ib} are the weights for each mixture component. According to Eq.(2), based on the weights of $\{\lambda_i\}$ generated from Step.6 of the previous section, interpolation distributions for EPU model i is:

$$P_i^{DIEPU}(\cdot) = \sum_a(\lambda_i w_{ia})\mathrm{N}(o;\mu_i,\sum_i) + \sum_b\{(1-\lambda_i)w_{ib}\}\mathrm{N}(o;\mu_i,\sum_i) \quad (4)$$

From Eq.(4), it is obviously that after sharing mixture components from well-trained models (IF) to refined but less well-trained models (EPU), the distribution of EPU can be smoothed, and with EM training, it will be optimized.

5. Experimental Results

The effectiveness of deleted interpolation approach was evaluated on spontaneous Mandarin speech using the LDC 1997 MBN corpus. The acoustic training set and interpolating set consisted of 10 hours of speech (10,483 utterances include about 183,513 syllables) selected from the first two CDs in the LDC 1997 MBN corpus. The 1997 MBN corpus was equally divided into 3 blocks, $M = 3$. The testing data consisted of two parts: the first test set (test_set1) included 865 spontaneous utterances with 11,512 syllables in total. Test_set1 was independent of the training set selected from the first two CDs. It included conversational speech, colloquial speech, the speech of people talking in a meeting, etc. The second test set (test_set2) was used for performance comparison consisting of clean utterances (F0 condition) from the 1997 and 1998 Hub4NE evaluation sets [9]. Test_set2 contained 1263 utterances, with about 15,535 syllables. The HMM topology was three-states, left-to-right without skips. The acoustic features were $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. 27 standard initials (include 6 zero initial symbols) and 38 finals were used to generate context-independent HMMs. We used decision tree based state-tying procedures to build 10 Gaussian-component triphone models with 2904 tied-states. 57 extended phone units were selected using DP alignment and confidence measure criterion. Based on acoustic and phonetic confusion measure criterion described in [10], we finally used 32 extended phone units which cover the majority of sound changes in spontaneous Mandarin speech. Among these 32 units, it is found that around 70% of them are for Chinese initials. The dictionary used in decoding was standard *syllable-to-initial/final* with multi-pronunciations, it had 2.4 pronunciations per syllable on average [9]. The total syllable numbers was

415. We combined interpolation weights learning method shown in Section 4 with EM training. In general, the interpolation weights converged after five or six iterations.

Using the decision tree based state-tying approach [11], 552 tied-states were generated for 96 decision trees of EPU triphones. The total number of Gaussians was 34,540 ((2904+552)*10). Compared with the baseline model of 29040 Gaussians, this gives a 18.9% increase in parameter size. In order to make a fair comparison, we generated an enhanced HMM which has 12 Gaussians per state (the total number of Gaussians is 34,848). We also compared the results with the method of Gaussian mixture sharing .The recognition performance is listed in Table 2.

Table 2: A comparison of recognition performance.

	Syllable Error Rate (SER) %	
system	**Test_set1**	**Test_set2**
Baseline	42.23	30.92
Enhanced HMMs	41.87	30.62
Gaussian mixture sharing	41.29	30.05
DI of EPU models	**40.25**	**29.78**

Table 2: A comparison of recognition performance.

It is shown that after interpolating some of EPU triphones, the SER reduces 1.98% and 1.62% absolutely on spontaneous speech compared with the baseline and the use of enhanced HMM, respectively. Furthermore, it gives an additional 1.04% absolute SER reduction in spontaneous speech with respect to that of Gaussian mixture sharing. It proves that after interpolating mixture components from well-trained models, the EPU models are smoothed and become more robust for speech recognition. The results in Table.2 also shows that simply increasing the Gaussian numbers per state does not help much in terms of SER reduction, since some of the Gaussians are poorly estimated as the number of Gaussians increased. In addition, the use of Gaussian mixture sharing only guarantee the shared mixtures with sufficient training samples can be estimated robustly.

6. Conclusion

We presented an approach of modeling sound changes in Mandarin spontaneous speech by deleted interpolating the sound-changed model with its corresponding baseline model. Based on standard phonetic unit, we used DP alignment together with data-driven method to extend the phone set automatically for sound change description. Our deleted interpolation approach was applied to mixture component weights rather than all HMM parameters, simplifying the procedure for continuous HMM. In addition, sound changes in Mandarin spontaneous speech are difficult to model because of the data sparseness problem, as well as it is very time consuming to generate hand-labeled transcriptions, our approach is still efficient to model sound changes at the phone and model levels. It has been shown that this deleted interpolation approach provides a significant 1.98% and 1.04% absolute SER reduction for baseline and Gaussian mixture sharing method in spontaneous speech.

7. References

[1] Frederick Jelinek, "Statistical Methods for Speech Recognition", the MIT Press, 1998.

[2] Y. Liu, P. Fung, "Modeling partial pronunciation variations for spontaneous Mandarin speech recognition", *Computer Speech and Language*, Vol.17, pp.357-379, 2003

[3] A. Li, X. Chen, G. Sun, et al, "The phonetic labeling on read and spontaneous discourse corpora," In *Proc. ICSLP2000*, Beijing, China, 2000.

[4] F. Zheng, et al., "Modeling pronunciation variation using context-dependent weighting and B/S refined acoustic modeling", In *Proc. Eurospeech2001*, Aalborg, Denmark, pp.57-60, 2001.

[5] M. Bacchiani and M. Ostendorf, "Joint acoustic unit design and lexicon generation," in *Proceedings of ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, Netherlands, pp. 7-12, 1998

[6] X.D. Huang et.al, "Deleted Interpolation and Density Sharing for Continuous Hidden Markov Models," In Proc. *ICASSP1996*.

[7] Nam Soo, Chong Kwan Un, "Statistically Reliable Deleted Interpolation," *IEEE Transactions on Speech and Audio Processing*, Vol.5, No.3 May 1997, pp292-295.

[8] M. Saraclar, H, Nock and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, 14,137-160, 2000.

[9] P. Fung, W. Byrne, F. Zheng, T. Kamm, Y. Liu, Z. Song, V. Venkataramani and U. Ruhi, Pronunciation modeling of Mandarin casual speech. Final report, The Johns Hopkins University Summer Workshop, 2000

[10] P. Fung and Y. Liu, "Effects and Modeling of Phonetic and Acoustic Confusions in Accented Speech Recognition". *Journal of the Acoustical Society of America*, Vol.118, Issue 5, pp.3279 – 3293, November 2005.

[11] S. Young, et al., *The HTK book (for HTK Version 3.2)*. Entropic Cambridge Research laboratory, 2002