interspeech 2008 incorporating SST 08

Phonetic Confusion Analysis and Robust Phone Set Generation for Shanghai-Accented Mandarin Speech Recognition

Guo-Hong Ding

Nokia Research Center, Beijing
guohong.ding@nokia.com

10.21437/Interspeech.2008-344

ABSTRACT

In this paper, accent issues are discussed for Shanghai-accented Mandarin speech recognition. The phonetic confusion is analyzed in detail based on the alignment between the surface form and the baseform transcriptions. Mutual information is used as the measure to extract the most confusing phoneme pairs. It was found that each phoneme in one pair can be easily misrecognized with the other. To remove the phonetic confusion, it is better to replace the two phonemes in one pair with a newly generated one. Consequentially new phone sets are derived. The phonetic confusion analysis and the experimental evaluation are performed on a Shanghai-accented Mandarin speech corpus. Experimental results show that compared to the canonical phone set, the generated one can reduce the substitution error greatly and achieve a 0.72% absolute Chinese character error rate (CER) reduction. When it is combined with pronunciation modeling, the absolute CER reduction is 1.58%.

1. INTRODUCTION

The recognition of native speech reaches in many cases an acceptable level. However, the processing of non-native accents remains among the most difficult tasks. Accent issue is getting crucial in automatic speech recognition with the broad applications of speech enabled services. When words or sentences are uttered with accented pronunciations, speech recognition performance is considerably degraded.

In China, there are 10 dialect families: Guan (Mandarin), Jin, Wu, Hui, Xiang, Gan, Kejia (Hakka), Yue (Cantonese), Min, and Ping. Althoug Putonghua (Mandarin) has been broadly chosen as the communicative spoken language, people with different dialectal backgrounds typically speak Mandarin with a certain degree of accent due to the influence of their mother tongue dialect. The influence could be phonetic, lexical, or syntactical [3].

Active research has been carried out on accented speech recognition during the past few years. The proposed methods vary from capturing the pronunciaiton variations present in accented speech or natural speech in a multiple pronunciaton lexicon using phonological knowledge or data-driven methods to compensate the acoustic model by acoustic manipulation or by acoustic adaptation. In [4], the context-dependent/context-independent Pinyin confusion was analyzed and applied for Shanghai-accented Mandarin speech recognition. In [2], besides the consideration of Pinyin confusion, standard speaker MLLR adaptation was applied to recognize Shanghai-accented speech using a recognizer trained from speech data collected in the Beijing area. In [6, 10], acoustic manipulation was performed at the state level to introduce accent information into native acoustic models. In [7], based on the investigation of the pronunciation variability between native and accented speech, the data of one confusing phoneme and its pronunciation variant were pooled together for decision tree-based state tying to train the model of the phoneme.

Obviously, all the above methods try to compensate pronunciation variation at the lexical level or at the acoustic level based on the confusion analysis between the accented speech and the canonical acoustic model, and promising results have been published. But the phonetic confusion of the standard phone set units on the accented speech itself is not considered yet. In other words, it is still open whether the standard phone set is suitable to provide a good balance between the demands of a high resolution acoustic model and the available accented data.

In this paper, we try to explore the phonetic confusion in accented speech and then to derive the robust phone set specific for Shanghai-accented speech. In the literature, robust phone set generation has been studied in spontaneous speech recognition tasks to tackle phonetic confusion problems. [1] proposed the Chi-square test to measure the confusion of the SAMPA-C phonetic units with short durations. In [5], the likelihood ratio test was used as the confidence measure in automatic phone set extensions to model phonetic confusion in spontaneous speech. Obviously, both [1] and [5] attempted to combine the standard phone set with new units to remove the phonetic confusion.

In this paper we make a different attempt. Based on the phonetic confusion analysis of Shanghai-accented speech, it was found that in the canonical phone set there exist some phoneme pairs and each phoneme is easily misrecognized as the other. Thus to remove phonetic confusion, it is better to replace the two phonemes in one confusing phoneme pair with a new phoneme to construct robust phone sets.

The rest of this paper is organized as follows. In Section 2, the accented speech corpus is described and the effect of the accent is explored briefly; Then in Section 3 the phonetic confusion is analyzed in detail and mutual information is used as the confidence measure to extract the most confusing phoneme pairs and consequentially new phone sets are generated; Experimental evaluation is provided in Section 5 to show the advantage of the extracted phone set compared to the original canonical one. Finally a brief conclusion is drawn in Section 6.

2. EFFECT OF SHANGHAI-ACCENTED SPEECH

In this section, a Shanghai-accented speech corpus is introduced and then the baseline recognition system is described. The effect of Shanghai-accented speech on the canonical phone set can be observed in the baseline experimental results.

2.1. Shanghai-accented speech and baseline ASR

The database used in this paper is RASC863 (Regional Accented Speech Corpus funded by National 863 Project), a Chinese speech corpus with 4 regional accents: Shanghai (Wu),

Guangzhou (Yue), Chongqing (Southwestern Mandarin) and Xiamen (Min) [3]. The corpus consists of spontaneous speech, read speech and selected dialectal words. All the speech data were recorded at 16 KHz using an earphone and a capacitor microphone at the same time.

In this paper, since the speech recognition task is large vocabualry continuous speech recogition for Shanghai-accented Mandarin, only the read speech data recorded in Shanghai with the capacitor microphone are used for analysis and experimental evaluation. In the read speech, 2200 phonetically rich sentences were automatically selected from the newspaper or the Internet on-line talk shows. The sentences cover all Chinese syllables, intersyllabic diphones, and 84% intersyllabic triphones. The sentences were divided into 20 sheets, and each speaker was assigned to read one sheet of 110 sentences. 200 speakers, balanced in terms of age, sex and educational background, were recruited for recording.

The Shanghai-accented Mandarin corpus mentioned above gives a good experimental basis to analyze the phonetic phenomena for accented speech. In the experimental setup, the recording script consists of 20 sheets, each of which was read by 10 speakers. It is better to assign one sheet (with 10 speakers) as the testing set, and others are the training set. Obviously, the speakers and the texts in the testing set are independent of those in the training set. Table 1 gives some necessary statistics on the two sets.

	Training set	Testing set
No. of Speakers	190	10
No. of Sentences	20981	1099
Total amount of Speech Data (hours)	49.44	2.24

Table 1. Statisitcs of training set and testing set

The 39-dimensional speech feature consists of 13 mel-frequency cepstral coefficients (MFCCs) including the logarithmic energy for every 10ms analysis frame, and their first and second derivatives. During training and testing, cepstral mean normalization and energy normalization are used before the extracted features are sent to the recognition engine. The acoustic model consists of 3-state context-dependent cross-word triphone HMMs, trained using the HTK toolkits [12]. The states of the triphone models are tied using the decision tree-based state clustering approach. Each state has 16 Gaussian mixtures with diagonal convariance matrices. As a result, 8186 triphones and 1227 states are obtained.

2.2. Effect of the accented speech

The canonical phone set used in this paper is modified from SAMPA-C [9]. It should be noted that for Pinyin “an”, the last letter “n” is subtly represented at the phonetic level as [n2] to differentiate the first letter “n” in Pinyin “na”.

Table 2 gives the free phoneme recognition results of the testing set on the baseline acoustic model. In the table, phoneme substituion is pretty serious. By analyzing the free phoneme reconigiton results, we could find that the probabilities for different phonemes being correctly recognized vary from 45.0% to 94.5%. Since the recognition is free from lexical constraints, obviously the misrecognition and the phoneme substitution are caused by the phonetic confusion when the canoncial phone set is used for the accented speech. Thus it is beneficial to analyze the phonetic confusion in detail and to explore good techniques to cope with the problem.

Sub.	Del.	Ins.	Err.
17.27	3.28	10.21	30.77

Table 2. Free phoneme recognition results of the testing set on the baseline acoustic model (%)

3. PHONETIC CONFUSION ANALYSIS AND ROBUST PHONE SET GENERATION

In this section, mutual information is used in confusion analysis between the canonical phone set and the training data and the most confusing phoneme pairs are extracted. By merging the confusing phoneme pairs as new phonemes, new robust phone sets are generated.

3.1. Examination of confusable phone set

The confusion relationship between phonemes X and Y is shown in Fig. 1. Two phonemes are assumed to be Gaussian variables. The threshold represents a measure criterion such as acoustic likelihoods and $p(S \mid B)$ denotes the probability of surface form S originated from baseform B. Obviously, in the dark area it is difficult to give an explicit measure of which baseform the surface form should be originated from, and consequentially phonetic confusion occurs.

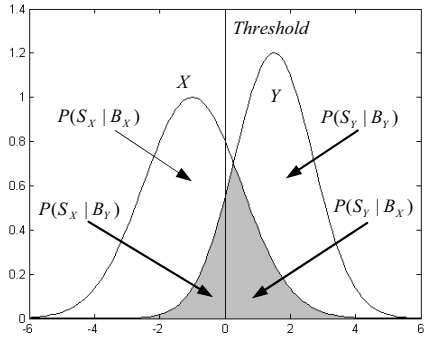

Fig. 1 Confusion relationship between phonemes X and Y

To derive $p(S \mid B)$, the alignment should be performed between the surface form and the baseform transcriptions. It has been described in detail in many papers [4, 5, 11]. First, the surface form transcription is obtained via free phoneme recognition of the training set on the baseline acoustic model. Then the baseform transcription is derived via forced alignment with the canonical lexicon. Finally, dynamic programming is used to align the two transcriptions and then the phonetic confusion matrix, which records the probabilities of one phoneme recognized as others, is extracted.

Our focus is on the characteristics of the canonical phone set in the crossed area for phonetic confusion analysis. Given the phonetic confusion matrix, it is interesting to analyze the most easily misrecognized units and the confusable pairs. Table 3 gives 12 phonemes with the lowest correctly recognized rates, while 12 phoneme pairs with the highest substitution error rates are listed in Table 4. In other words, the 12 lowest diagonal units of the confusion matrix are listed in Table 3, while the highest off-diagonal units are listed in Table 4.

From Table 3, it is obvious that the most easily mis-recognized phonemes are the consonants [ts_h, ts`, ts`_h, ts, s, z`, s`], the phonetic realization of the initials [c, zh, ch, z, s, r, sh], respectively and the vowels [i`, @`, o, {_, i\], which

correspond to finals [i, er, o, a, i], respectively. According to the phonetic definition [9], vowels [i`, i\] are the different phonetic representations of letter [i] in different contexts. Table 4 shows that consonants [s, ts`, ts_h, ts, ts`_h, s`, N, n2] are easily misrecognized as [s`, ts, ts`_h, ts`, ts_h, s, n2, N], respectively, while the vowels [{_, i`, i\, y] may possibly be recognized as [E_, i\, i`, i]. Obviously, there exist some pairs, in which one phoneme is easily misrecognized as the other.

Phoneme	ts_h	ts`	ts`_h	ts	s	i`
Corr.	60.46	62.57	63.86	67.10	67.47	67.92
Phoneme	@`	o	z`	{_	s`	i\
Corr.	68.05	70.31	71.91	72.67	75.04	76.36

Table 3. 12 lowest diagonal units of the confusion matrix (%)

Base.	s	ts`	ts_h	{_	ts	ts`_h
Sur.	s`	ts	ts`_h	E_	ts`	ts_h
Prob.	19.46	16.04	15.84	15.68	15.46	15.26
Base.	i`	s`	N	n2	i\	y
Sur.	i\	s	n2	N	i`	i
Prob.	14.25	13.44	12.49	12.39	12.38	11.88

Table 4. 12 highest off-diagonal units of the confusion matrix (%)

3.2. Pronunciation modeling

Since the surface form and the baseform transcriptions have been extracted, the pronunciation modeling can be done directly as described in [11]. It should be noted that in [11] the pronunciation modeling tries to cope with spontaneous speech and it can be extended simply to accented speech. The extracted pronucniation lexicon will be used in the experimental evaluation later.

3.3. Phonetic confusion analysis using mutual information

Mutual information measures the information different variables share. In other words, it measures how much knowledge one variable reduces the uncertainty about the other. The mutual information of surface form *S* and baseform *B* is defined as

$$I(S:B)=\sum_{s\in S}\sum_{b\in B}p(s,b)\log\frac{p(s,b)}{p(s)p(b)}$$

As a result, to measure the confusion between two different phonemes *X* and *Y*, it is beneficial to analyze the dependence of one phoneme on the other using the cross mutual information (CMI). For the two phonemes, the cross mutual information can be formulated as $I(S_X:B_Y)$ to represent the dependence of surface form S_X on baseform B_Y or $I(S_Y:B_X)$ to represent the dependence of S_Y on B_X.

Table 5 lists 12 phoneme mappings with the highest cross mutual information. In comparison with Table 4, obviously the cross mutual information can give a different dependence measure compared to conditional probabilities. Besides, the phoneme mapping from baseform [{_] to surface form [E_] listed in Table 4 is missing in Table 5. It should be noted that though the conditional probability of surface form [E_] given baseform[{_] is pretty high, the occurrence of [{_] is much less compared to other phonemes (the marginal probability of the baseform [{_] is 0.27%). As a result, the phonetic confusion contributed by this phoneme mapping will give less influence on the total phonetic confusion and consequentially bring about less impact on recognition performance.

Base.	s`	ts`	i`	n2	N	ts`_h
Sur.	s	ts	i\	N	n2	ts_h
CMI	8.53	7.29	7.00	6.43	5.15	4.56
Base.	ts	s	ts_h	i\	y	ts\
Sur.	ts`	s`	ts`_h	i`	i	ts_h
CMI	4.45	3.33	2.52	2.09	1.68	1.60

Table 5. 12 phoneme mappings with highest cross mutual information ($\times 0.001$)

3.4. Robust phone set generation

To measure the inter-dependence (confusion) of two different phonemes, it is better to explore the total cross mutual information (TCMI), namely, the sum of $I(S_X:B_Y)$ and $I(S_Y:B_X)$. The total cross mutual information of the most confusable phoneme pairs is listed in Table 6.

Original phonemes	TCMI($\times 0.01$)	Merged phoneme
s,s`	1.19	s_sh
ts,ts`	1.17	ts_sh
n2,N	1.16	N_sh
i`,i\	9.09	i_sh
ts_h, ts`_h	7.07	ts_h_sh
z`,l	2.82	

Table 6. 6 confusable phoneme pairs with the highest total cross mutual information and possible phoneme merges

From Table 6, the total cross mutual information of the top 5 pairs, namely, (s,s`), (ts,ts`), (n2,N), (i`,i\), and (ts_h,ts`_h), are much higher than others and obviously they are the most confusable phoneme pairs in Shanghai-accented speech. It is interesting to analyze the phonetic characteristics of the confusable phoneme pairs. For (s, s`), (ts, ts`) and (ts_h, ts`_h), the only difference is that one phoneme is dentoalveolar, while the other is retroflex For (i`, i\), since the former always appears after [s`, ts`, ts`_h] and the latter after [s, ts, ts_h]. Obviously, the confusion more or less comes from the confusion of the context. For (n2, N), both are the phonetic representation of letter "n" and the only difference is that [n2] is dentoalveolar nasal, while [N] is velar nasal.

Besides, it is observed from Table 5 that one phoneme in each of the 5 pairs may be easily misrecognized as the other and vice versa. Thus it is a problem how to differentiate one phoneme from the other in one pair. One solution is to cancatenate one phoneme in the confusing pairs with other phonemes to form a new longer phoneme to improve the robust capability [1]. In this paper, we attempt to simply merge the two phonemes in one pair as a new phoneme as it is shown in Table 6. That means, for each pair the new phoneme will replace the two phonemes to construct a new phone set.

Since the TCMIs are different for the 5 confusing phoneme pairs, it is better to extract 5 phone sets according to different confusion degrees. In other words, in PhoneSet1 only [s, s`] are merged, in PhoneSet2, both [s, s`] and [ts, ts`] are merged, and ...

4. EXPERIMENTAL EVALUATION

In this section, speech recognition experiments are performed to evaluate the performance of the generated phone sets for free phoneme recognition and word recognition.

The baseline acoustic model training has been described in Section 2. To train the model and to perform word recognition on the new phone sets, the training transcription, the question set and the lexicon should be modified to accommodate the new phonemes. Since in the generated phone sets some original canonical phonemes are simply replaced by new phonemes, it is easy to get the new training transcriptions and the new lexicons. As to the question set, as it was analyzed in Section 3.4, the two phonemes in each pair have limited phonological difference, and as a result only the conflicting questions need to be ignored from the original question set.

	Sub	Del	Ins	Err
Baseline	17.27	3.28	10.21	30.77
PhoneSet1	16.56	3.25	10.01	29.81
PhoneSet2	15.87	3.21	9.96	29.03
PhoneSet3	14.27	3.45	9.79	27.51
PhoneSet4	13.95	3.45	9.54	26.95
PhoneSet5	13.44	3.39	9.60	26.43

Table 7. Free phoneme recognition results for different modified phone set along with the baseline results (%)

The free phoneme recognition results are listed in Table 7. In the free phoneme recognition experiments, no language models are used and the recognition is purely performed by acoustic matching. For each phone set, the results are obtained by matching the recongized phoneme series with the phoneme scripts, which are generated by aligning the word scripts with the corresponding acoustic model and the phone set. From the table, it can be concluded that with the increase of merged phonemes, the substitution rates are reduced consistently, while the deletion and the insertion rates are varied slightly. For PhoneSet4, the absolute phoneme substitution rate reduction is 3.32% and the absolute phoneme error rate reduction is 3.82%.

Baseline	29.35
PhoneSet1	29.16
+PM	28.23
PhoneSet2	28.87
+PM	28.30
PhoneSet3	28.76
+PM	28.14
PhoneSet4	28.63
+PM	27.77
PhoneSet5	28.83
+PM	28.01

Table 8. Chinese character error rates for different phone sets plus pronunciation modeling along with the baseline results (%)

Table 8 gives the word recognition results. It should be noted that in the word recognition evaluation, both the original lexicon and the pronunciation lexicon mentioned in Section 3.2 are used to give the phonetic representations for Chinese words. In the word recognition experiments a bigram word-based languge model is used. The language model is originally trained using SMS text data [8] and adapted on the training scripts of the Shanghai-accent Mandarin corpus. Since the SMS text data are totally different from the testing scripts, which mainly consist of news, the perplexity is pretty high. But it is still valuable to evaluate the extracted phone sets in word recognition with the language model. In the table it is found that all the generated phone sets are superior to the baseline and the pronunciation modeling (PM) can provide further performance improvement. For PhoneSet4, the absolute reduction of Chinese character errors is 0.72%, and with the introduction of pronunciation modeling, the absolute reduction is 1.58%.

5. CONCLUSION AND FUTURE WORK

In this paper, accent issues are explored for Shanghai-accented Mandarin speech recognition. The phonetic confusion analysis is based on the alignment between the baseform and the surface form transcriptions. Mutual information is used to extract the confusing phoneme pairs, in which each phoneme may be easily misrecognized as the other. To generate robust phone sets, it is better to merge the two phonemes in one pair as a new one. Experimental results show the advantage of the generated phone sets.

Robust phone set generation is a good attempt for accented speech and experimental evaluation shows promising results. But it still remains open on how to utilize the accent-dependent phoneme confusion along with the accented speech data in one Mandarin recognizer with the canonical acoustic model to cope with multiple accents.

6. REFERENCES

[1] Y.-J. Chen, et al., Generation of Robust Phonetic Set and Decision Tree for Mandarin using Chi-Square Testing, *Speech Communication*, 38: 349-364, 2002.

[2] C. Huang, E. Chang, J. Zhou, K.-F. Lee, Accent Modeling Based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition, *In Proc. ICSLP*, 2000.

[3] A. Li, Z. Yin, T. Wang, Q. Fang, F. Hu, RASC863 – A Chinese Speech Corpus with Four Regional Accents (http://ling.cass.cn/yuyin/report/files/2004_15.pdf), *In Proc. Oriental-COCOSDA*, 2004.

[4] M. Liu, B. Xu, T. Huang, Y. Deng, C. Li, Mandarin Accent Adaptation Based on Context-Independent Context-Dependent Pronunciation Modeling, *In Proc. ICASSP*, 2000.

[5] Y. Liu, P. Fung, Automatic Phone Set Extension with Confidence Measure for Spontaneous Speech, *In Proc. Eurospeech*, 2003.

[6] Y. Liu, F. Zheng, L. He, Y. Xia, State-Dependent Mixture Tying with Variable Codebook Size for Accented Speech Recognition, *In Proc. ASRU*, 2007.

[7] Y. R. Oh, J. S. Yoon, H. K. Kim, Acoustic Model Adaptation based on Pronunciation Variability Analysis for Non-Native Speech Recognition, *Speech Communication*, 49: 59-70, 2007.

[8] J. Olsen, Y. Cao, G. Ding, X. Yang, A Decoder for Large Vocabulary Continuous Short Message Dictation on Embedded Devices, *In Proc. ICASSP*, 2008.

[9] SAMPA-C, http://ling.cass.cn/yuyin/english/sampac/sampac.htm.

[10] M. Saraclar, H. Nock, S. Khudanpur, Pronunciation Modeling by Sharing Gaussian Densities across Phonetic Models, *Computer Speech and Language*, 14: 137-160, 1999.

[11] M.-Y. Tsai, F.-C. Chou, L.-S. Lee, Pronunciation Modeling with Reduced Confusion for Mandarin Chinese Using a Three-Stage Framework, *IEEE Trans. Audio, Speech and Language Proc.*, 15(2): 661-675, 2007.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.0*. Cambridge, U.K.: Cambridge Univ. Press, 2000.