Acoustic and Phonetic Confusions in Accented Speech Recognition

Yi Liu and Pascale Fung

Department of Electrical and Electronic Engineering
University of Science and Technology, Hong Kong
(eeyliu, pascale)@ee.ust.hk

10.21437/Interspeech.2005-147

Abstract

Accented speech recognition is more challenging than standard speech recognition due to the effects of phonetic and acoustic confusions. Phonetic confusion in accented speech occurs when an expected phone is pronounced as a different one, which leads to erroneous recognition. Acoustic confusion occurs when the pronounced phone is found to lie acoustically between two baseform models and can be equally recognized as either one. We propose that it is necessary to analyze and model these confusions separately in order to improve accented speech recognition without degrading standard speech recognition. We propose using likelihood ratio test to measure phonetic confusion, and asymmetric acoustic distance to measure acoustic confusion. Only accent-specific phonetic units with low acoustic confusion are used in an augmented pronunciation dictionary, while phonetic models with high acoustic confusion are reconstructed using decision tree merging. Experimental results show that our approach is effective and superior to methods modeling phonetic confusion or acoustic confusion alone in accented speech, with a significant 5.7% absolute WER reduction, without degrading standard speech recognition.

1. Introduction

Most state-of-the-art automatic speech recognition (ASR) systems fail to perform well when the speaker has a regional accent different from that of the standard language the systems trained on. The high error rate is partly due to the effects of phonetic confusions and acoustic confusions in accented speech recognition. Phonetic confusion in accented speech is caused by the speaker pronouncing an expected phone in a different way (for example, when /zh/ is pronounced as /z/). On the other hand, acoustic confusion arises when the accented speech is found to lie acoustically somewhere between two baseform phones and can be equally recognized as either (for example, when it is in between /zh/ and /z/). Acoustic confusion can also come from data and recognizer-related confusions, in addition to pronunciation variation [1, 2]. Previous studies on accented speech recognition focused on detailed effects of either phonetic confusions or acoustic confusions, but not both at the same time. Extending the phonetic unit set [3] and generating accent-specific dictionary with multiple pronunciations [4] are commonly used for modeling phonetic confusions. To model acoustic confusions, one can either improve the discriminative ability of acoustic models [5] or use adaptation to modify the acoustic parameters [4].

However, the difference and correlation between phonetic and acoustic confusions, and their different roles in ASR systems are less clear. For example, a poorly trained acoustic model with acoustic confusions leads to high phonetic confusions since the speech recognizer is a pattern recognition task that is primarily based on acoustic distance measure. Similarly, phonetic confusions can lead to acoustic confusions due to the erroneous phonetic labels in acoustic model training. Therefore, it is essential to distinguish phonetic and acoustic confusions in accented speech in order to achieve better recognition performance.

In this paper, we propose methods to measure phonetic and acoustic confusions and reduce them for optimal speech recognition performance on accented speech without sacrificing the performance on standard speech.

The paper is organized as follows. Section 2 outlines the difference and correlations of phonetic and acoustic confusions in accented speech. In Section 3, we describe the approach of reducing phonetic and acoustic confusions in accented speech recognition. In Section 4, experiments on accented Mandarin telephony speech are presented. We conclude in Section 5.

2. Acoustic confusions vs. phonetic confusions

2.1. Acoustic and phonetic confusions are different yet correlated

In accented speech, phonetic confusion is caused by the pronunciation of an expected phone into a different one whereas acoustic confusion arises from a pronounced phone lying between two standard phones acoustically. For a speech recognizer trained on standard speech, phonetic confusion is then the erroneous recognition of a phonetic unit in the accented speech into another phonetic unit in the standard speech. It can be regarded as the probability of the transformation from a baseform unit to a surface form unit. Acoustic confusion, on the other hand, is at a more fundamental level and describes the distance between the phonetic unit in accented speech and phonetic units represented by two baseform models, in terms of acoustic properties.

Phonetic and acoustic confusions are different yet correlated in speech recognition task. If the acoustic models of two phonetic unites are close to each other (i.e. not easily separable), then these models have low discriminative ability and will cause phonetic confusions in the final recognition task, irregardless of whether the input speech is accented or not. However, even if the trained acoustic models have good

separability, accented speech might produce a phone that lies somewhere between two models and again cause acoustic confusion, resulting in phonetic confusion. In other cases, the accented speech might produce one phone that is clearly close to another, different phone in the standard speech. This causes phonetic confusion, even though there is no acoustic confusion between models.

We use Fig.1 to illustrate the distinction and correlation between phonetic confusion and acoustic confusion. Suppose 'A' is a phonetic unit and 'B' is another phonetic unit that is often confused with 'A'. Acoustic models for 'A' and 'B' consist of a single Gaussian component, $G_A(\mu_A, \sigma_A)$ and $G_B(\mu_B, \sigma_B)$, respectively, where μ and σ are the mean and the variance. The phonetic confusion between units 'A' and 'B' can be measured using $P(B \mid A)$ that can be computed from the number of occurrence [1]. The more mapping pairs of 'A' and 'B' they are, the higher the phonetic confusion is. On the other hand, the acoustic confusion is measured using the acoustic distance between models 'A' and 'B', i.e., the distance between the Gaussian components. The more model 'A' and model 'B' are overlapped (the shaded area in Fig.1), the higher the acoustic confusion between 'A' and 'B' is. Obviously, this type of confusion is measured differently from the previous one.

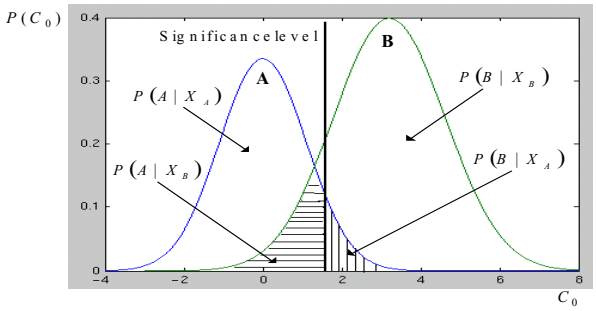

Fig.1: Relations in phonetic and acoustic confusions.

2.2. Measuring phonetic and acoustic confusions

Phonetic confusions are measured in terms of the distribution of the mapping between surface form and baseform phones. We use *likelihood ratio test* to evaluate the degree of phonetic confusions to take into account of the original distribution of the phonetic unit. The log of the likelihood ratio λ is as follows:

$$\log\lambda = \log L(c_{12}, c_1, p) + \log L(c_2 - c_{12}, N - c_1, p) - \log L(c_{12}, c_1, p_1) + \log L(c_2 - c_{12}, N - c_1, p_2), \quad (1)$$

where $L(k, n, x) = x^k (1-x)^{n-k}$ is a binomial distribution. In Eq. (1), c_1, c_2 and c_{12} are the total number of occurrences of the a canonical phonetic unit, the confusing unit and the aligned confusing pair of their combination. p, p_1 and p_2 can be calculated as

$$p = \frac{c_2}{N}, \; p_1 = \frac{c_{12}}{c_1}, \; p_2 = \frac{c_2 - c_{12}}{N - c_1}, \quad (2)$$

where N is the total number of phonetic units in the training set.

The degree of acoustic confusions can be measured by the dissimilarity or distance between two speech vectors, between a speech vector and a speech model, and between two speech models. For accented speech, we are interested in measuring the statistical dissimilarity between that of accented speech model and standard speech model. We propose an *asymmetric acoustic distance measure* to measure the degree of acoustic confusions. The acoustic distance is inversely proportional to acoustic confusion. Suppose λ_i and λ_j are two different CHMM phonetic models that consist of N states. Each individual state is represented by a probability distribution function (PDF) in terms of multiple Gaussian components. s_{in} and s_{jn} are two different states of model λ_i and model λ_j, the acoustic distance between λ_i and λ_j is expressed as

$$D(\lambda_i, \lambda_j) = \sum_{n=1}^{N} D(s_{in}, s_{jn}), \quad (3)$$

where

$$D(s_{in}, s_{jn}) = \min_{W=[w_{kl}]} \sum_{k=1}^{K} \sum_{l=1}^{L} w_{kl} d\left(g_{in,k}, g_{jn,l}\right), \quad (4)$$

where $d\left(g_{in,k}, g_{jn,l}\right)$ is the distance between two single Gaussian components; $W = [w_{kl}]$ is the weight matrix and can be determined using Simlex tableau method [6]; K and L are total Gaussian mixture numbers of each state. We use asymmetric form of Mahalanobis distance for element distance measure in which only σ_2 is used

$$d\left(g_{in,k}, g_{jn,l}\right) = \left(\mu_k - \mu_l\right)^T \sigma_2^{-1} \left(\mu_k - \mu_l\right). \quad (5)$$

This type of asymmetric form is in accordance with the asymmetric and unidirectional properties of acoustic confusions in accented speech.

2.3. Combinations of phonetic and acoustic confusions

There are four combinations of acoustic and phonetic confusions in speech recognition systems: (1) phonetic confusions and acoustic confusions are both low; (2) phonetic confusion is low and acoustic confusion is high; (3) phonetic confusion is high and acoustic confusion is low; and (4) phonetic confusions and acoustic confusions are both high.

Ideally, the subword units (e.g., phonemes and phones or initials/finals in Mandarin speech) used in ASR systems should be modeled and trained so that phonetic and acoustic confusions are both low for good discriminative-ness [2]. Condition (1) is therefore desirable for ASR systems.

Condition (2) in which phonetic confusion is low but acoustic confusion is high is relatively rare. It happens when two phoneme models are acoustically confusable (i.e. with overlapping acoustic characteristics such as between /l/ and /n/), but accented speaker tend to distinguish the two phones very clearly, even more so than standard speakers (for example, Cantonese speakers never pronounce /l/ close to /n/). This type

of confusion exists for accented speech where native models are acoustically confusable (e.g. “l” and “n” in Mandarin) whereas accented speakers, by over compensation, can separate the two pronunciations better than native speakers in their pronunciation (e.g. Cantonese speakers of Mandarin) [7]. Under condition (2), accented speech does not adversely affect speech recognition performance.

Condition (3) under which phonetic confusion is high and acoustic confusion is low indicates that phonetic confusion in this case is not caused by acoustic confusion, since acoustic models under this condition have good discriminative abilities. Accented speech is a predominant factor leading to phonetic confusion in this case. For instance, acoustic confusion between models /f/ and /x/ is low since there is little overlapping acoustic characteristic between standard Mandarin models of these two sounds.

Under condition (4), phonetic and acoustic confusions are both high. If most of the phonetic units are phonetically and acoustically confused, then perhaps the unit inventory is not well defined and/or the acoustic models are not well trained. The acoustic models do not have good separability and ASR performance will suffer greatly. Another factor is again accent. In most cases, the two factors co-exist. That is, the acoustic models do not have good separability *and* the accented speech differs from standard speech.

3. Reducing ponetic and acoustic confusions for accented speech recognition

We studied the four combinations of acoustic confusions and phonetic confusions in accented speech recognition. The investigation of these four combinations and the corresponding pronunciation phenomena in accented speech shows that phonetic and acoustic confusions should be considered separately to improve recognition performance in accented speech recognition task. Fig. 2 gives examples of acoustic and phonetic distances of Chinese initials in the accent-specific units.

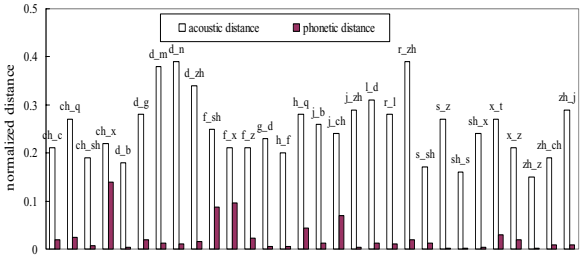

Fig.2: Examples of normalized acoustic and phonetic distances in extended units.

To model phonetic and acoustic confusions in accented speech for the task of speech recognition, we propose the following algorithm for modeling phonetic and acoustic confusions in accented speech:

1. Identify phonetic confusion in the input speech by *likelihood ratio test* to generate a set of units $\{ P \}$;
   1a. Identify accent-specific confusion pairs from $\{ P \}$ by using dialectical pronunciation rules, and replace set $\{ P \}$ by this new set $\{ P_{accent} \}$.
2. Identify acoustic confusion from $\{ P_{accent} \}$ using *asymmetric acoustic distance measure*, and form a set of units that have high phonetic confusion but low acoustic confusion $\{ A\_l \}$ and another set of units with high phonetic confusion as well as high acoustic confusion $\{ A\_h \}$;
3. For phonetic units in $\{ A\_l \}$, form a multiple pronunciation dictionary with *extended phone set*;
4. For phonetic units in $\{ A\_h \}$, use *acoustic model reconstruction* with decision-tree merging.

State-transition charts of the above algorithms are shown in Fig. 3. Having classified accent-specific phonetic units according to high and low acoustic confusions, we suggest selecting only phonetic units with low acoustic confusions to form alternate pronunciations and add to a pronunciation dictionary. For phone units with high acoustic confusions, we suggest that incorporating them into a pronunciation dictionary will further lead to lexical confusions. Instead, we propose using decision tree merging with acoustic model reconstruction for this class of phone units. More details about forming alternate pronunciations with augmentation of dictionary and decision tree merging of acoustic model reconstruction can be found in our previous work [5].

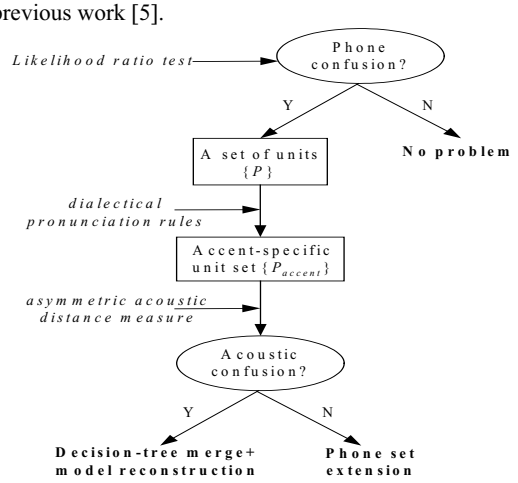

Fig. 3: State-transition charts of modeling phonetic and acoustic confusions in accented speech.

4. Experiments

We use a Mandarin telephony speech database in our experiments. All speech data were sampled with 8 kHz and 8 Bit. The baseform acoustic model was trained using 50 hours of native Mandarin speech. 2,000 continuous utterances with 23,685 syllables from 20 Cantonese accented speakers (DATA1) were used to extract the accent-specific units. The acoustic features are $13MFCC$, $13\Delta MFCC$ and $13\Delta\Delta MFCC$. We used HTK decision tree based state tying procedures to build 12

Gaussian-component triphone models with 5500 tied-states. The testing data contains two parts. Both test_set1 and test_set2 include nine speakers (4 females and 5 males), and each speaker has 100 phrase utterances. Utterances in test_set1 and test_set2 are the same. Speakers in test_set1 are all Cantonese accented while in test_set2 are standard Mandarin speakers.

Using DATA1 as the development set, we obtained 79 accent-specific units with high phonetic confusions. When applying the asymmetric acoustic distance measure, only 22 units with low acoustic confusions are selected to form alternative pronunciations and generated a multiple pronunciation dictionary with high phonetic confusion, low acoustic confusion units (Dict3). Comparative results from using a standard dictionary and a dictionary (Dict2) reweighted and augmented using all phonetically confusable units [8] are shown in Table 1. We can see that the use of Dict3 yields additional 1.1% absolute WER reduction on accented speech compared to using Dict2. In addition, there is no adverse impact on standard speech. The improvement comes from modeling phonetic confusions without introducing acoustic confusions and lexical confusions in Dict3.

System	Word Error Rate (WER) %	
	(Test_set1)	(Test_set2)
Baseline Standard phone set (Dict1)	20%	7.9%
Phone set augmented by all high phonetic confusion units (Dict2)	18% (-2.0)	7.7% (-0.2)
Phone set augmented by only phonetically confusable units with low acoustic confusions (Dict3)	**16.9% (-3.1)**	**7.7% (-0.2)**

Table 1: A comparison of WER of using different multiple pronunciation dictionaries with respect to the baseline.

The use of multiple pronunciation dictionary with selected extended units is able to model phonetic confusions but not to acoustic confusions. Based on our previous work, we use the approach of acoustic model reconstruction to refine the pre-trained acoustic model and to achieve a high discriminative ability for reducing acoustic confusions, together with the use of the Dict3, we modeled both phonetic confusions and acoustic confusions. The results are shown in Table 2. It is seen that the approach of acoustic model reconstruction is able to improve the resolution of acoustic model with high discriminative ability for modeling acoustic confusions, and yields 4.8% absolute WER reduction. Combined with the use of multiple pronunciation dictionary based on selected extended units, we can model both acoustic and phonetic confusions. It gives 5.7% WER reduction compared with the baseline. In addition, our method provides a better WER reduction on accented speech recognition without sacrificing the performance on native speech. Our approach can be applied to a single system to handle both accented and native speech, and even speech with multiple accents.

5. Conclusions

We study the effects of phonetic confusions and acoustic confusions in accented speech. We suggest that phonetic and acoustic confusions are different yet correlated in accented speech. We suggest that only phone units which lead to high phonetic confusions in accented speech cause recognition errors. Among these units, there are those that also have high acoustic confusions and others with low acoustic confusions. We propose to model these two classes of phone units differently for better recognition performance on both accented and standard speech. We use likelihood ratio test to select units with high phonetic confusions and we propose an asymmetric acoustic distance measure to describe the unidirectional properties of acoustic confusions in accented speech. Experimental results in Cantonese accented Mandarin speech show that the combination of modeling phonetic confusions and acoustic confusions yields a 5.7% absolute WER reduction, compared to 3.1% of modeling phonetic confusion alone, and to 4.8% of modeling acoustic confusion alone. Meanwhile, there is no performance degradation on standard speech recognition.

6. References

[1] M. Saraclar, et al., "Pronunciation modeling by sharing Gaussian densities across phonetic models", *Computer Speech and Language,* (2000) 14, 137-160

[2] M.Y. Tsai, et al., "Pronunciation variation analysis based on acoustic and phonetic distance measure with application examples on Mandarin Chinese", Proc. *ASRU03*, pp.117-121, 2003

[3] Y.J. Chen, et al., "Generation of robust phonetic set and decision tree for Mandarin using chi-square testing", *Speech Communication* 38, 349-364

[4] Ch. Huang, et al., "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition", Proc. *ICSLP00*, 2000

[5] Y. Liu, et al., "Partial change accent models for accented Mandarin speech recognition", Proc. *ASRU03*, 2003

[6] Z. Liu, et al., "A new distance measure for probability distribution function of mixture types", Proc. *ICASSP00*, pp. 1345-1348, 2000

[7] X. Huang et al., *Chinese Dialects*, Xia Men University Press, 1987 (Chinese version)

[8] M. Riley, et al., "Stochastic pronunciation modeling from hand-labeled phonetic corpora", *Speech Communication* 29, 209-224

System	Word Error Rate (WER) %	
	(Test_set1)	(Test_set2)
Baseline	20%	7.9%
Multiple pronunciation dictionary (modeling phonetic confusion alone)	16.9% (-3.1)	7.7% (-0.2)
Reconstructed HMMs (modeling acoustic confusion)	15.2% (-4.8)	7.1% (-0.8)
Reconstructed HMMs & selected multiple pronunciation dictionary	**14.3% (-5.7)**	**7.1% (-0.8)**

Table 2: Our approach outperforms the baseline, and modeling phonetic or acoustic confusion alone.