PRONUNCIATION VARIATION ANALYSIS BASED ON ACOUSTIC AND PHONEMIC DISTANCE MEASURES WITH APPLICATION EXAMPLES ON MANDARIN CHINESE

Ming-Yi Tsai and Lin-Shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan, Republic of China

pancho@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this paper, two conceptually different statistical distance metrics are defined and analyzed. First, the asymmetric *acoustic distance* measures how the acoustic property of one phoneme is close to that of another, and is defined here based on the Mahalanobis distance between two hidden Markov models. Second, the asymmetric *phonemic distance* measures how probable a phoneme is realized as another, based on the aligned phonemic canonical and surface transcriptions of speech corpora. These two mutually dependent distance measures are used to construct an abstract *acoustic/phonemic distance* plane which is helpful in quantitatively analyzing pronunciation variations. Besides, clear distinction and complicated correlation between these two distances were discussed, which is believed to be helpful in many application areas involving lexical design and acoustic modelling as well. Preliminary analysis were performed on LDC Hub-4NE Mandarin Broadcast News database, and possible application in pronunciation modelling was discussed.

1. INTRODUCTION

Statistical distance measures between two hidden Markov models have been applied in various areas of speech recognition, for example the acoustic model estimation [1], phoneme model clustering [2], deriving sub-word units [3], mapping multilingual phoneme units [4, 5], speaker clustering [6], vocabulary selection [7, 8], or unit selection in concatenative speech synthesis [9]. However, there is still few applications of statistical distance measures for pronunciation variation analysis. In this paper, instead of taking the uniform acoustic distance between phonemes from base-form and surface form [12], we measure the confusability among phonemes by means of two conceptually different statistical distance metrics, namely the *acoustic distance* and the *phonemic distance* to investigate both the acoustic and phonemic confusion which will be confronted during recognition with the same phoneme inventory. Moreover, as will be clear later on, instead of using the knowledge-based "*phonetic distances*" obtained by vectors of articulatory features [10], the statistical data-driven metrics as developed in this paper are not only capable of taking into account the real acoustic characteristics of speech signals (for example, it is well known that the differentiation between two phonetically different consonants by either human or machine can be seriously affected by its acoustic environment, such as the channel distortion or masking effects), but also applicable to all languages and application tasks as long as a corpus for analysis is available,

without the requirements for linguistic knowledge.

By defining the two statistical distance measures, the *acoustic distance* (i.e. how the acoustic property of a phoneme is close to that of another) and the *phonemic distance* (i.e. how probable a phoneme is realized as another), it will be clear below that the distinction and correlation between these two distance measures are very helpful in analyzing pronunciation variation. For example, considering a speech production process [11], when a speaker produces a word, he first selects an appropriate item from his mental lexicon (i.e. the base form), and then the articulatory shape of this item is prepared (i.e. surface form). The *phonemic distance* here is an asymmetric measure regarding how probable the transformation from the base form to its surface form of one phoneme would take place. And such *phonemic distance* (degree of *phonemic confusion*) is usually represented in a form of a recognition dictionary for explicit pronunciation modelling. On the other hand, the speech recognition process is primarily based on the *acoustic distances* among acoustic models trained on acoustic signals while incorporating the *phonemic distance* or degree of *phonemic confusion* represented in the recognition dictionary with explicit pronunciation modeling. From a modeling point of view, a good recognition dictionary should try to include more *phonemic confusion* (i.e. the transformation from a base form to its surface form) but avoid more *acoustic confusion* (i.e. phoneme pairs which the machine couldn't distinguish well due to small *acoustic distance*) at the same time.

The organization of this paper is as follows. The *acoustic* and *phonemic distance* measures to be investigated are first introduced in Section 2, and the corpus and framework for the analysis are then presented in Section 3. In Section 4, the preliminary analysis for the pronunciation variation of Mandarin Chinese based on these distance measures is discussed. Further considerations regarding the approach of retraining acoustic models commonly used in pronunciation modelling are mentioned in Section 5, and the conclusion is finally given in Section 6.

2. ACOUSTIC AND PHONEMIC DISTANCE MEASURES

In this paper, two conceptually different distance measures are utilized in analyzing pronunciation variation. One is the *acoustic distance*, or the difference in acoustic properties from one phoneme to another, based on the statistical parameters of the corresponding HMMs. Here we use the asymmetric form of Mahalanobis distance to achieve this purpose. The other is an asymmetric *phonemic distance*, or how probable a base form phoneme is transformed

to a surface form, based on statistics obtained from the phonemic alignment of canonical and surface transcriptions of speech corpus.

2.1. Acoustic Distance from One N-state Multi-mixture Hidden Markov Model (HMM) to Another

The asymmetric form of Mahalanobis distance is used in this paper for measuring the differences in acoustic properties, due to its relatively simple formulation and its capabilities in the circumstance of asymmetric acoustic confusion, although it is not known at this moment if some other distance measures may serve the purpose better.

For two Gaussian distributions g_1 and g_2 , with mean vectors μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 respectively, the asymmetric form of Mahalanobis distance is defined as [13]

$$d(g_1, g_2) = (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2) \quad (1)$$

which is obviously asymmetric because only Σ_2 is used.

Considering two N -state M -mixture Gaussian hidden Markov models, λ_i and λ_j , the asymmetric form of Mahalanobis distance is then defined below, which is the *acoustic distance* used in this paper,

$$D_{ac}(\lambda_i, \lambda_j) = \sum_{s=1}^N \sum_{m_{i,s}=1}^M w_{m_{i,s}} \sum_{m_{j,s}=1}^M w_{m_{j,s}} d(g_{m_{i,s}}, g_{m_{j,s}}) \quad (2)$$

where $w_{m_{i,s}}$ is the mixture weight of the $m_{i,s}$ -th Gaussian distribution $g_{m_{i,s}}$ in state s of model λ_i , and so on. Based on this definition, the asymmetric *acoustic distance* $D_{ac}(\lambda_i, \lambda_j)$ can be comprehended as the averaged distance of a sample signal generated by λ_i from model λ_j , and $D_{ac}(\lambda_j, \lambda_i)$ as the distance of a sample signal generated by λ_j from model λ_i . So it describes the differences in acoustic properties from one HMM to the other, and it is asymmetric.

2.2. Phonemic Distance from One Phoneme to Another

Here the *phonemic distance* is to measure how probable one base-form phoneme is transformed to another phoneme in surface form, or the “*phonemic confusion*” from one phoneme to another. Obviously, the phonemic confusion is asymmetric and the definition used in this paper is different from the symmetric form which was used in the literature for vocabulary selection [7]. Furthermore, we wish to distinguish the phonemic distance from the acoustic distance defined above.

The proposed phonemic distance between two phonemes, u_i and u_j , is therefore defined here in this paper as

$$D_{ph}(u_i, u_j) = -\ln[Pr(T_s = u_j | T_c = u_i)] \quad (3)$$

where T_s and T_c are the aligned phonemic surface and canonical transcriptions respectively. Apparently the more frequently u_i is pronounced as u_j , the smaller the phonemic distance $D_{ph}(u_i, u_j)$ is, and this distance measure can be evaluated based on the empirical statistics obtained from the phonemic alignment of canonical and surface transcriptions of speech corpus.

Fig. 1. An abstract acoustic/phonemic distance plane defined by the phonemic and acoustic distances.

2.3. Analysis with an Acoustic/Phonemic Distance Plane

In order to analyze the pronunciation variation better with the two different distance measures defined above, an abstract *acoustic/phonemic distance* plane is illustrated in Fig.1. In this plane, the horizontal scale stands for the *phonemic distance* (the more to the right the larger the *phonemic distance* is, or the more to the left the more phonemically confused one phoneme is to another). For example, if a pair of phonemes are located in the right half plane (region I and region II) in Fig.1, one phoneme is in general hardly pronounced as the other. On the other hand, since the acoustic and phonemic distances are certainly mutual-dependent (not orthogonal), the scale of *acoustic distance* is not absolutely vertical in this plane. Similarly, the lower in the plane a pair of phonemes are located, one is more confused with the other in acoustic properties, and so on. Therefore, the distance plane can be more or less divided into four regions as shown in Fig.1, which is very helpful in analyzing the pronunciation variation. For example, most pairs of phonemes in a language should belong to region I (*phonemic* and *acoustic distance* are both high), so they can be easily distinguished. However, phoneme pairs in region II (with low *phonemic distance* but high *acoustic distance*) usually correlate with those relatively frequent transformations from one phoneme to another phonetically quite different phoneme (e.g. with quite different place or manner of articulation), mainly due to the pronunciation convention or assimilation processes. Besides, in region III (*phonemic* and *acoustic distance* are both low), pairs usually correlate with transformations of two phonetically similar phonemes. On the other hand, pairs in region IV (high *phonemic* but low *acoustic distance*), usually correlate with two phonemes which are phonetically similar, but people rarely pronounce one as the other.

It has been well acknowledged that the surface form transcriptions, whether it is generated automatically by machine or manually by human, always include some errors. How to derive correct phonological transformations from such erroneous surface form transcriptions by identifying those errors has long been a difficult problem. The abstract *acoustic/phonemic distance* plane analysis presented here may offer some directions in handling this problem. For example, the transformation pairs in region II (low *phonemic* but high *acoustic distance*), are more likely indicate correct phonological transformations, but more transcription pairs in region III (both low *phonemic* and *acoustic distance*), are possibly produced by errors in the surface form transcription.

Fig. 2. Phonemic pronunciation variation statistics of Mandarin INITIALs, where *Freq* is the percentage of unit frequency in corpus, *P_c* is the probability of canonical pronunciation, and *R_{del}* is the deletion rate of each INITIAL respectively.

Place	Manner	Liquid	Nasal	Aspirated Stop	Unaspirated Stop	Aspirated Affricative	Unaspirated Affricative	Unaspirated Fricative	Aspirated Fricative
Bilabial			m	p	p'				
Labiodental								f	
Dental Sibilant						dz	ts	s	
Dental and Alveolar			n	d	d'				
Retroflex						dz'	ts'	s'	ʒ
Alveolo-palatal						dz\	ts\	s\	
Velar				g	k			h	

Table 1. Mandarin INITIALs with respect to the place and manner of articulation.

3. CORPUS AND FRAMEWORK FOR SOME EXAMPLE ANALYSIS

Traditional Chinese phonology divides the a Mandarin syllable into two sub-syllabic units, an INITIAL and a FINAL. The INITIAL, or onset, is the way a syllable begins, usually with a consonant. FINAL. The INITIAL, or onset, is the way a syllable begins, usually with a consonant. A small number of syllables, such as /an/ and /a/, do not begin with a consonant. They are said to begin with the zero INITIAL, which is labelled as /#/ in this paper. The FINAL of a syllable is the syllable minus the INITIAL. The longest form of a FINAL consists of three parts: a medial; a main vowel, or nucleus; and an ending, or coda. There may not be a medial or ending in some syllables, but there must be at least a main vowel in each syllable[14].

The pronunciation variation of Mandarin INITIALs is analyzed on the LDC Hub-4NE 1997 Mandarin Broadcast News database. About 30 hours of speech is used to train gender-dependent and context-independent acoustic models, consisting of 22 three-state INITIALs and 37 four-state FINALs regardless of the tone. The tone are temporarily ignored here. There are 24 Gaussian mixtures per state. The acoustic features are 13 MFCCs, 13 delta MFCCs and 13 acceleration MFCCs. The gender-dependent acoustic models trained based on the canonical transcription, is then used to automatically transcribe the phonemic surface forms of the same 30 hours of speech data. To avoid the "noisy" recognition errors, a refined automatically transcribing procedure [15] is used. The broad phonetic transcription used here is based on *SAMPA-T* [16] through this paper.

Through our analysis, we found that the phonemic pronunciation distribution patterns are similar in male and female speech. However, the degree of variation is more significant in female than in male speech. Therefore, we will only present the analysis results on female speech. Those for male speech and FINALs are

not included in this paper simply due to the limited space.

4. PRONUNCIATION VARIATION IN MANDARIN INITIALS

The phonemic pronunciation variation statistics of Mandarin INITIALs in Hub-4NE corpus is shown in Fig.2. The vertical bar and the two curves in Fig.2 represent respectively the unit frequency (*Freq*), the probability of canonical pronunciation (*P_c*), and the deletion rate (*R_{del}*) of each INITIAL in the corpus. And the INITIALs are sorted with increasing probability of canonical pronunciation (*P_c*). From Fig.2, it is clear that the degree of pronunciation variation (as represented by *P_c*) is apparently less correlated with the unit frequency (as represented by *Freq*). Besides, it can be found from the curve of *P_c* in Fig.2 that /n/, /p/, /t/, /t'/ are the four most non-canonical-pronounced INITIALs, and /m/, /l/, /b/ and /t/ are the next four most non-canonically-pronounced INITIALs, while all other INITIALs are more canonically pronounced. This situation is roughly demonstrated with three gray levels in Table.1, in which the 21 Mandarin INITIALs, except for zero INITIAL /#/ , are listed with respect to their place and manner of articulation. A systematic pattern can be found in Table.1 that the bilabial, labiodental and dental-alveolar (except for /d/) INITIALs are more likely to be realized in alternative forms. In other words, the pronunciation variation for Mandarin INITIALs seems to be more correlated with the place rather than the manner of articulation. Also, it can be found from the *R_{del}* values in Fig.2 that the INITIALs /t'/, /#/ , /n/, /m/, /l/ are more likely to be deleted in pronunciation. Checking with Table.1, it seems to be more correlated with the manner of articulation as well. In addition, all voiced (or semivowel) INITIALs are found to be less canonically pronounced as well.

The *acoustic* and *phonemic distances* among all Mandarin INITIALs, calculated with metrics defined in Eqs.(2) and (3), are

Fig. 3. Acoustic distance among Mandarin INITIALS presented in the order of their degree of variation. The smaller the distance, the darker the indexed block is.

shown respectively as two matrices in Figs.3 and 4. The labels of INITIALS for the matrices are in the same order of their degree of pronunciation variation as those in Fig.2. The distance (or degree of confusion) is represented by the brightness in the matrices, i.e. the smaller the distance from a INITIAL to another the darker the indexed block. Note that the distances here are asymmetric. The INITIALS listed in the vertical scale of the matrix stands for the first units in Eqs.(2) and (3), while those on the horizontal scale stands for the second units. For example, in the *acoustic distance* matrix in Fig.3, the brightness pattern of a row may be considered as the averaged distance pattern for sample signals of one INITIAL, labelled on the left end of the row, to all INITIALS, labelled on the horizontal scale on the top of the matrix. Similarly, in the *phonemic distance* matrix in Fig.4, the brightness pattern of a row stands for the phonemic distance pattern of one INITIAL, labelled on the left end of the row, to all INITIALS, labelled on the horizontal scale on the top of the matrix.

From the *acoustic distance* matrix in Fig.3, it can be found that the 8 most non-canonically-pronounced INITIALS, /n/, /p/, /t/, /Z'/, /m/, /l/, /b/ and /f/, on the top-left 8x8 square in the matrix, are acoustically very confused with all INITIALS, in particular with the 8 INITIALS themselves. That's why the top-left 8x8 square is very dark, and the top 8 rows and left 8 columns of the matrix are all relatively dark. In addition, the asymmetric property is also clear here. For example, /Z'/ is correlated with the darkest column (vertical bar) but not the darkest row (horizontal bar) in the matrix. This implies that /Z'/ is acoustically not the closest to all other INITIALS (as implied by its row), but it has relatively higher acoustical variety (or larger $\sum_2$ in Eq.(2), as implied by its column). This should come from the inherent characteristics of its pronunciation characteristics. Furthermore, in the matrix in Fig.3, the four voiced Mandarin INITIALS, /l/, /m/, /n/ and /Z'/, are found symmetrically very close to each other in *acoustic distance*, but this is not true for the *phonemic distance* in Fig.4. For example, as observed in Fig.4, /n/ is not phonemically confused with, or hardly pronounced as /Z'/. Besides, in Fig.4, /l/, /m/ and /n/ are phonemically confused with the zero INITIAL /#/, but the *acoustic*

Fig. 4. Phonemic distance among Mandarin INITIALS presented in the order of their degree of variation. The smaller the distance, the darker the indexed block is.

	-iEn		-in		-an		-a		-@N		-ei
#	0.56	#	0.42	d	0.39	d	0.38	d	0.54	n	0.33
n	0.21	n	0.18	n	0.29	n	0.21	n	0.23	d	0.23
l	0.05	l	0.11	l	0.14	#	0.15	#	0.11	#	0.20

Table 2. Pronunciation confusion of Mandarin INITIAL /n/ to a few most frequent realizations conditioned on the right context in Hub-4NE corpus

distances as shown in Fig.3 do not imply this. By comparing the matrices in Figs.3 and 4, it is clear that the acoustic and phonemic distance measures are quite different and both asymmetric, as mentioned previously.

Some interesting phenomenon were observed from the *phonemic distance* matrix in Fig.4. Firstly, the "bright region" in the lower-left part of the matrix implies that those usually canonically-pronounced INITIALS (with higher P_c and at the lower part in the vertical scale of the matrix) are rarely pronounced as those usually non-canonically-pronounced INITIALS (with lower P_c and at the left part in the horizontal scale of the matrix). In addition, the darkness distribution is more or less uniform and relatively darker over the entire right half matrix. This implies that almost all INITIALS tend to be pronounced as those usually canonically-pronounced INITIALS (with higher P_c and at the right half in the horizontal scale of the matrix). In other words, when most INITIALS, in particular those often canonically pronounced, are pronounced in alternative way, they tend to be realized as those usually canonically-pronounced INITIALS. Second, it was also found in Fig.4 that INITIAL /n/ is very often realized as the zero INITIAL /#/, /d/ and /l/, but the *acoustic distance* from /n/ to /#/ and /d/ are much larger (the not-so-dark blocks in the first row of Fig.3). Clearly, the transformation from /n/ to /#/, /d/ and /l/ are all typical assimilation processes. Some examples of the assimilation process on /n/ given its right context in Hub-4NE corpus are listed with the corresponding probabilities in Table.2 as a reference.

The above discussions lead to the concept of the abstract *acoustic/phonemic distance plane* as illustrated in Fig.1 previ-

Fig. 5. Acoustic and phonemic distance values from /n/ to all Mandarin INITIALs in Hub-4NE corpus (from the first rows in both Fig.3 and Fig.4).

Fig. 6. Acoustic and phonemic distance values from /d/ to all Mandarin INITIALs in Hub-4NE (from the first rows in both Fig.3 and Fig.4).

ously, in which the four regions can be used to identify different nature of phoneme transformations. We picked up the most non-canonically pronounced INITIAL /n/, and an usually canonically pronounced INITIAL /d/ as two examples. The *acoustic/phonemic distance* from /n/ and /d/ to all INITIALs in Hub-4NE corpus are depicted in Figs.5 and 6 respectively, which are in fact respectively the distance values in the first rows of Figs.3 and 4 for /n/, and in the third row from the bottom of Figs.3 and 4 for /d/. With Figs.5 and 6 it is easy to compare the values of the two distances, and to locate transformation pairs from /n/ and /d/ to all other INITIALs on the four regions of Fig.1. Several typical examples of such transformations are shown in Fig.7. As can be seen from Fig.7, the assimilation process from /n/ to /#/ and from /n/ to /d/ mentioned above are in region II, while that from /n/ to /l/ is in region III, etc. All discussions regarding the four regions mentioned previously then apply here.

5. ACOUSTIC DISTANCE MEASURES FOR RETRAINED ACOUSTIC MODELS

Conventionally, acoustic models are trained with the canonical transcription of speech data. However such "canonically trained" acoustic models may be less "purely trained" because the training data may include different realizations. This may degrade the machine recognition accuracy, and make the understanding on pronunciation variation more difficult. In order to improve the machine recognition performance, many researchers tried to retrain the acoustic models based on some "selective" surface form tran-

Fig. 7. Typical examples of several transformation pairs from /n/ and /d/ located on the abstract acoustic/phonemic distance plane.

scriptions. This was also done here with the present experiment. The acoustic distance matrix among Mandarin INITIALs, just as that in Fig.3, was also obtained on the retrained acoustic models, which is shown in Fig.8. Comparing Fig.8 with Fig.3, almost no noticeable difference is found. This may explain why very limited, or even negative improvements on recognition accuracy was achieved by retraining acoustic models in literatures [17]. Since the machine recognition is primarily based on the acoustic distance, the recognition accuracy won't be improved if the acoustic distance among the units remain unchanged. This indicate that such distance measures may provide an alternative approach in analyzing machine recognition performance.

6. CONCLUSION AND FUTURE WORK

In this paper, a new approach of analyzing pronunciation variation based on *acoustic* and *phonemic distance* measures is presented. An abstract *acoustic/phonemic distance* plane was constructed by the conceptually different *acoustic* and *phonemic distances*. It not only successfully provides the distinct description between *acoustic* and *phonemic* confusion, but also reveals their interdependency.

In addition, by analyzing with the *phonemic distance*, it was found that when most INITIALS, in particular those often canonically-pronounced, are pronounced alternatively, they tend to be realized as other often canonically-pronounced INITIALS. On the other hand, with the *acoustic distance*, it was found that those most non-canonically-pronounced INITIALS, /n/, /p/, /t/, /tʰ/, /m/, /l/, /b/ and /f/, are more acoustically confusable with all INITIALS. Furthermore, it was also observed that the degree of pronunciation variation is less correlated with the unit frequency in speech data. And phonemic pronunciation variation seems to be more correlated with the place than the manner of articulation for Mandarin INITIALS.

Both the distinction of *acoustic/phonemic confusion* and the pronunciation variation distribution developed in this work are very helpful especially from a pronunciation modelling perspective. In the future, we will improve the Human-Machine Communication Model for simulating the word correct rate to compile a pronunciation variation dictionary [11] by including more *phonemic confusion* in the speech production module and avoid possible *acoustic confusion* in the transmitting channel at the same time, based on development from this paper.

Fig. 8. Acoustic distance measures among Mandarin INITIALS for retrained acoustic models, presented in the order of their degree of variation. The smaller the distance, the darker the indexed block is.

7. REFERENCES

- [1] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden markov models," *ATT Technical Journal*, vol. 64 No.2, pp. 391–408, 1985.
- [2] Matti Vihola, Mikko Harju, Petri Salmela, Janne Suontausta, and Janne Savola, "Two dissimilarity measures for hmms and their application in phoneme model clustering," in *ICASSP*, 2002.
- [3] Rita Singh, Bhiksha Raj, and Richard M. Stern, "Structured redefinition of sound units by merging and splitting for improved speech recognition," in *ICSLP*, 2000.
- [4] Joachim Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *ICSLP*, 1996.
- [5] A.K. Kienappel, D. Geller, and R. Bippus, "Cross-language transfer of multilingual phoneme models," in *ISCA ITRW ASR*, 2000.
- [6] Zhipeng Zhang and Sadaoki Furui, "An online incremental speaker adaptation method using speaker-clustered initial models," in *ICSLP*, 2000.
- [7] P. Geutner, M. Finke, and A. Waibel, "Selection criteria for hypothesis driven lexical adaptation," in *ICASSP*, 1999.
- [8] Beng T. Tan, Yong Gu, and Trevor Thomas, "Word confusability measures for vocabulary selection in speech recognition," in *ASRU*, 1999.
- [9] Jon Yi and James Glass, "Information-theoretic criteria for unit selection synthesis," in *ICSLP*, 2002.
- [10] Eric Fosler-Lussier and Nelson Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, pp. 137–158, 1999.
- [11] Ming-Yi Tsai and Lin-Shan Lee, "Pronunciation modeling for spontaneous speech by maximizing word correct rate in a production recognition model," in *The ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003.
- [12] Murat Saraclar and Sanjeev Khudanpur, "Pronunciation ambiguity vs pronunciation variability in speech recognition," in *ICASSP*, 2000.
- [13] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, 1973.
- [14] Yuen-Ren Chao, *A grammar of spoken Chinese*, vol. 1, University of California Press, Berkeley and Los Angeles, 1965.
- [15] Ming-Yi Tsai, Fu-Chiang Chou, and Lin-Shan Lee, "Improved pronunciation modeling by properly integrating better approaches for baseform generation, ranking and pruning," in *ISCA Workshop: Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, Estes Park, Colorado USA, 2002.
- [16] Chiu-yu Tseng and Fu-chiang Chou, "Machine readable phonetic transcription system for chinese dialects spoken in taiwan," *J. Acoust. Soc. Jpn*, vol. 20, pp. 215–223, 1999.
- [17] Michael Riley, William Byrne, Michael Finke, Sanjeev Khudanpur, Andrej Ljolje, John McDonough, Harriet Nock, Murat Saraclar, Charles Wooters, and George Zavaliagkos, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, pp. 209–224, 1999.