

AUTOMATIC GENERATION OF PRONUNCIATION LEXICONS FOR MANDARIN SPONTANEOUS SPEECH

*W. Byrne*¹, *V. Venkataramani*¹, *T. Kamm*¹, *ZHENG F.*², *SONG Z.*², *P. Fung*³, *LIU Y.*³, *U. Ruhi*⁴

CLSP/ECE, The Johns Hopkins University, Baltimore MD, USA (1)
Dept. EEE, Hong Kong University of Science and Technology, Hong Kong (3)

Dept. CST, Tsinghua University, Beijing, China (2)
Dept. CS, University of Toronto, Canada (4)

ABSTRACT

Pronunciation modeling for large vocabulary speech recognition attempts to improve recognition accuracy by identifying and modeling pronunciations that are not in the ASR systems pronunciation lexicon. Pronunciation variability in spontaneous Mandarin is studied using the newly created CASS corpus of phonetically annotated spontaneous speech. Pronunciation modeling techniques developed in English are applied to this corpus to train pronunciation models when are then applied in Mandarin Broadcast News transcription.

1. INTRODUCTION

Pronunciation modeling for large vocabulary speech recognition attempts to improve recognition accuracy by identifying and modeling pronunciations that are not in the ASR systems pronunciation lexicon. These novel pronunciations are observed in spontaneous or casual speech, in accented speech, or due to coarticulatory effects not indicated in the lexicon. In this work we focus on pronunciation variation in spontaneous Mandarin speech. We make use of the Chinese Annotated Spontaneous Speech Corpus (CASS) [1], which is a newly created corpus of closely annotated, spontaneous Mandarin speech. Pronunciation models trained on the CASS corpus are refined and applied in the transcription of Mandarin Broadcast News (MBN).

We employ the decision-tree based pronunciation modeling methodology that has proven effective for English read speech and conversational speech recognition [2, 3, 4]. This approach casts pronunciation modeling as a prediction problem. The goal is to predict the variations in the surface form (i.e. in the phonetic transcription provided by expert transcribers) given the baseform pronunciation derived from a pronunciation lexicon and a word transcription. We use the CASS transcriptions as the expert annotations required by this approach.

The speech in the CASS corpus was provided by the Broadcast Station of Tsinghua University, Beijing, China and consists primarily of impromptu addresses, delivered in an informal style without prompts or written aids. The collection contains a total of three hours of speech, spoken by two female and five male adults. All speakers have lived for several decades in Beijing, a Mandarin speaking city.

This work was supported by the National Science Foundation under Grant No. #IIS-00712125, and carried out at the 2000 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Johns Hopkins University.

The detailed annotation in CASS consists of transcriptions at the word level, syllable level, and semi-syllable level with detailed pronunciation variants. The semi-syllable transcriptions use the SAMPA-C [5] convention, which is a computer readable alternative to IPA, specialized for Chinese.

Given the relatively small amount of phonetically transcribed data available for training it is not possible to obtain reliable estimates for all pronunciation phenomena: only events that occur often in the corpus can be modeled well. A conservative approach in this situation would model only those pronunciation changes observed often in the annotated data. A more ambitious approach which has proven effective would be to build large models using the annotated data, accept that they are poorly estimated, and then refine them using additional, acoustic data. This additional data is transcribed lexically, but not phonetically. This refinement process applies the initial models to the transcriptions of the acoustic training set, and an existing set of acoustic models is used to select likely pronunciation alternatives by forced alignment [2, 3, 4]. The refinement effectively discards spurious pronunciations generated by the initial set of trees and also 'tunes' the pronunciation models to the acoustic models that will be used in recognition.

In the pronunciation modeling work upon which this project is based [2, 3, 4], phonetic annotations were available within the ASR task domain of interest. This allowed decision trees to be refined using speech and acoustic models from the same domain as the data used in building the initial models. While it would be ideal to always have annotated data within the domain of interest, it would be unfortunate if new phonetic transcriptions were required for all tasks. However, unlike acoustic models and language models, pronunciation lexicons have been proven to be effective across domains. This suggests that, apart from words and phrases that might be specific to particular task domains, it is reasonable to expect that pronunciation variability is also largely domain independent.

We take the optimistic view that an initial set of pronunciation models trained on the CASS SAMPA-C transcriptions will generalize well enough so that they contain pronunciation variability in the Mandarin Broadcast News domain. We will identify which of the broad collection of alternatives inferred from the CASS domain actually occur in the Mandarin Broadcast News domain. After this refinement of the alternatives, new pronunciation lexicons and acoustic models will be derived for the Mandarin Broadcast News domain. In summary, we will take advantage of the refinement step to adapt CASS pronunciation models to a Broadcast News ASR system.

2. NORMALIZATION AND ACOUSTIC ANNOTATION OF THE CASS TRANSCRIPTIONS

The SAMPA-C annotation plays two roles in the CASS database. Most frequently, the SAMPA-C annotation describes the speech of Mandarin speakers more closely than would be possible using Pinyin. But it can also describe extreme deviation from the defined. In this latter situation, SAMPA-C symbols indicate sounds that never occur in any standard pronunciation. As a result, the CASS SAMPA-C transcriptions contain annotated initials that are not in the pronunciation lexicon used to train the Broadcast News baseline system. Even within the CASS domain itself, these symbols are not found in the dominant pronunciation of any word. It is therefore not clear under what circumstances these variants should be preferred over pronunciations containing the standard phones and, given this uncertainty, it is difficult to train acoustic models for these symbols. Rather than develop a suitable training method for these unusual initials, we replaced them by their nearest ‘standard’ initials. In the CASS transcription, initials marked as voiced p_{-v} , t_{-v} , k_{-v} , c_{-v} , ch_{-v} , q_{-v} , sh_{-v} , and s_{-v} were replaced by b , d , g , z , zh , j , zh , and z , respectively. We note that this substitution is not as exact as the transcription suggests, since Mandarin initials are not usually voiced.

2.1. Direct Measurement of Predictive Features

The use of annotation that indicates the present or absence of articulatory features suggests the possibility of directly measuring the features in the acoustic signal, by detecting voicing, nasalization and/or aspiration. For example, given the speech “bang fu”, a voicing detector could be used to produce the automatically annotated form “b_{-v} ang_{-v} f_{-v} u_{-v}”, indicating contextual voicing effects on the f . In this way, the pronunciations derived by lexicon from a word transcription can be augmented by direct acoustic measurements of the speech. These measurements could be used to aid in the prediction, i.e. to provide additional side information to improve the choice of pronunciation by the ASR system.

Voicing is relatively easy to measure directly, although much voicing information is admittedly represented by the per-frame energy measurements in the cepstral acoustic features; direct measurement of voicing may therefore not provide much additional information. However, we incorporate voicing measurements into our pronunciation models. Using the Entropic *get_f0* pitch tracking program [12], a frame-by-frame voicing decision for the entire CASS corpus was computed (frames were 7.5msec with a 10msec step). The time segmentation in the SAMPA-C tier allows us to transform these per-frame decisions into segment-based probabilities of voicing, P_v , by counting the number of voiced frames and dividing by the total number of frames in the segment.

It was observed in CASS corpus that 60% of the data is marked as voiced. We therefore identified the following procedure for normalizing the voicing detection: for each speaker, find the speaker-dependent *voice_bias* threshold in *get_f0* such that 60% of the segments had $P_v > 0.5$. The parameter value that gave the closest match to the 60% voiced criteria was chosen. The performance of this voicing detection scheme can be measured by comparison to voicing information inferred from the CASS transcriptions. We found an equal error point of 20% miss and 20% false alarm.

2.2. Introducing Variability into the CASS Transcriptions

The SAMPA-C tier of the CASS corpus is a very accurate transcription of the acoustic data. To non-expert listeners, however, these transcriptions seem very similar to dictionary-derived transcriptions. The consensus among native Chinese speakers is that the expert phoneticians can find evidence for sounds by listening to the speech and studying spectrograms that casual listeners can not find. This raises difficulties for our use of HMM acoustic models with these transcriptions. It is reasonable to question whether these acoustic models will be able to identify pronunciation changes that native (albeit inexpert) listeners cannot detect. As expected, many of the initials and finals that casual listeners would prefer to delete are quite short according to the time segmentation in the SAMPA-C tier. This poses another difficulty for our acoustic models, which are not well suited for modeling very short acoustic segments.

We addressed these difficulties by discarding the shortest initials and finals in the CASS transcriptions when building the pronunciation models. This introduced variability by removing the initials and finals from the transcription that we reasoned would be the most difficult to hear and to model. Through these modifications to the original CASS transcriptions, disagreement between the canonical and surface form pronunciations increased from 2.2% to 11.3%. It may seem that discarding data in this way is fairly drastic. In fact, despite this aggressive intervention the canonical pronunciations remain the dominant variant. Furthermore, these transcriptions are used only to train the pronunciation models; they are not used to train acoustic models. In the acoustic alignment steps that use these models, the acoustic models will be allowed to choose alternative forms, if it leads to a more likely alignment than that of the canonical pronunciation.

3. MANDARIN ACOUSTIC EQUIVALENCE CLASSES

The use of acoustic classes for constructing decision trees for pronunciation modeling has been discussed at length in [2, 3, 4, 6]. We adopt the approach as developed for English, although some alterations are needed in its application to Mandarin. In English the phonetic representation is by individual phones and acoustic classes can be obtained relatively simply from IPA tables, for instance. In Mandarin, word pronunciations are conventionally given in Pinyin, which specifies the syllables and tones that make up a word. The syllable inventory is fixed: our dictionary contains 403 unique syllables, disregarding tone. Each syllable has a standard pronunciation. The subsyllable units used in our system are initials and finals, which are an optional initial consonant followed by the remainder of the syllable. Because of this, the acoustic classes of the finals cannot be determined from the acoustic features of any individual phone. The classes were constructed instead with respect to the most prominent features of the initials and finals, and fall into the following broad categories; we also use explicit voicing notation, since voicing change is the most frequent effect observed in CASS.

Manner of Articulation of Finals. Identify with open vowels, “stretched” vowels, retroflex vowels, and protruded vowels. Finals of type ending in the nasals n and ng are also distinguished.

Place of Articulation of Finals. Identify high front, central, front, middle, and back. Front vowel finals ending in n and ng are also distinguished, and *uang* is distinguished by itself.

Vowel Content of Finals. Identify the number of vowels in their canonical pronunciation. Finals are also distinguished if they

end in *n* and *ng*, and the retroflex final is separated from all others.

Manner of Articulation of Initials. Identify aspirated and unaspirated stops, aspirated and unaspirated fricatives, nasals, voiced and unvoiced fricatives, and laterals.

Place of Articulation of Initials. Identify as labial, alveolar, dental sibilants, retroflex, dorsal, and velar.

Main Vowel of Finals. Identify finals sharing a main vowel.

Vowel Content and Manner of Finals. Identify monophones, diphthongs, and triphthongs as open, rounded, stretched, or protruded.

4. MANDARIN BROADCAST NEWS TRANSCRIPTION

The baseline MBN system was trained using a 50,614 word dictionary containing Pinyin pronunciations of words and individual characters; the baseline system was toneless. The acoustic training set consisted of 10 hours of speech (10,483 utterances) selected from the first two CDs in the LDC Mandarin 1997 Broadcast News distribution. MFCC acoustic features were used.

The word/character dictionary was used to segment the training set transcriptions so that pronunciations were available for the entire training transcription. Baseline word, syllable (toneless Pinyin), and initial/final transcriptions were derived for the training set using this segmentation and the baseline dictionary.

Context dependent initial/final models were trained using Baum Welch training and acoustic clustering procedures [9]. The HMM topology was left-to-right, without skips; models used for initials had three states, while models used for finals had four states. The HTK flat-start procedure was used to build 12-mixture Gaussian, state clustered HMMs with 2086 states.

The SRI language modeling tools [10] was used to train a bigram language model using news text resegmented to agree with the baseline dictionary. The following text was used: People's Daily, 1978-1996 : 233M words (approx); China Radio International, scripts : 56M words (approx); Xinhua newswire text 13.2M words (approx). The total corpus contained 303.2M words and the resulting bigram contained 50,624 unigrams and 7,992,589 bigrams.

4.1. MBN Pronunciation Models from CASS Transcriptions

The objective is to augment the baseline dictionary with pronunciation alternatives inferred from the CASS transcriptions. These new pronunciations are word-internal, in that pronunciation effects do not span word boundaries. The steps in the training procedure are as follows.

1. CASS Initial/Final Transcriptions The baseform initial/final transcriptions were derived from the Pinyin tier in the modified CASS transcriptions, and the surface forms were taken from CASS SAMPA-C tier. The initial and final entries were tagged by the voicing detector, using the time marks in the CASS SAMPA-C tier, as described.

2. CASS Decision Tree Pronunciation Models The transcriptions form the training set used to construct the initial decision tree pronunciation models to predict variations in the CASS data. A separate decision tree is trained for each initial and final in the lexicon. The trees are grown under a minimum entropy criterion that measures the purity of the distribution at the leaf nodes; cross-validation is also performed over 10 subsets of the training corpus to refine the tree sizes and leaf distributions [3, 4]. Changes in each initial or final can be predicted by asking questions about

the preceding 3 and following 3 symbols; questions about neighboring surface forms were not used.

3. Mandarin Broadcast News Pronunciation Lattices The CASS decision tree pronunciation models were applied to the MBN acoustic training set transcriptions, producing for each utterance a lattice of pronunciation alternatives; this step yields pronunciation alternatives for utterances in the Broadcast News training set based on alternatives learned from the CASS transcriptions. Segments in the MBN transcriptions were tagged with voicing information via forced alignment using the baseline MBN system.

3. CASS Decision Tree Pronunciation Alternatives The most likely pronunciation alternatives for the MBN training set were found by forced alignment through the pronunciation lattices. The alignment used two-component Gaussian mixture monophone HMMs from the baseline system to avoid using more complex models that might have been 'overly exposed' to the baseform transcriptions in training.

4. CASS Decision Tree Word Pronunciations The forced alignments yield alternative pronunciations for each entire utterance. In these initial experiments we wish to have only the pronunciation alternatives for individual words. We found these by first: aligning the words in the transcription with align the surface form phonetic transcription; and then tabulating the frequency of every alternative pronunciation for each word. The result of this step is a *Reweighted and Augmented CASS Dictionary* tuned to the baseline Broadcast News acoustic Models.

5. Broadcast News Decision Tree Pronunciation Models The alignments between the surface form and baseform sequences over the MBN training set are used to train another set of decision tree pronunciation models.

6. MBN Decision Tree Word Pronunciations The decision tree pronunciation models obtained in the previous step can be applied to the Broadcast News acoustic training transcriptions to produce for each utterance a trellis of pronunciation alternatives; these alternatives are a refinement of the first set of alternatives that were chosen from CASS pronunciation models.

7. State-Level Surface Form to Base Form Alignments An alternative Viterbi alignment can be performed to obtain an alignment at the state level between the surface form pronunciations and the baseform pronunciations; this alignment uses the fully trained MBN acoustic models. From this it can be determined which states in the baseform transcription are 'confusable' with the surface form transcription obtained under the pronunciation model. The most confusable states can be considered as candidates in soft-state clustering schemes [13, 6]. For each state in the HMM system, we found the most confusable state in the surface form paths, and a copy of each Gaussian from this state is added to the baseform mixture distribution; several Baum Welch iterations are performed to update the means and mixture weights in the entire system. This is the first HMM re-training step in the modeling procedure; all preceding steps are concerned with finding alternate pronunciations and refining the estimates of their frequencies.

8. Most Likely Word Pronunciations Under the Broadcast News Decision Tree Pronunciation Models A second *Reweighted and Augmented Broadcast News Dictionary* can be found in a manner identical to the first, except that the surface form pronunciations are chosen from trellises generated by the Broadcast News decision tree pronunciation models. A new dictionary is then constructed to include the most frequent word pronunciation alternatives.

These training steps generate two *Reweighted and Augmented*

Min. Count / Min. Relative Frequency	New Pronunciations	%CER and Improvement
Baseline	0	28.7
3 / 0.05	1664	28.4 (0.3)
3 / 0.1	1640	28.3 (0.4)
3 / 0.4	1490	28.4 (0.3)
0 / 0.01	6001	28.8 (-0.1)
0 / 0.05	5890	28.4 (0.2)
0 / 0.1	5739	28.3 (0.4) *
0 / 0.4	4490	28.3 (0.4)

Table 1: Performance of Reweighted and Augmented CASS Dictionary on the MBN Test Set with Baseline Acoustic Models.

Dictionary	%CER and Improvement
Baseline	27.8 (0.9)
Reweighted/Augmented CASS *	27.5 (1.2)

Table 2: Performance of Soft Reclustering of HMM States. The CASS dictionary used is marked by * in Table 1.

Dictionaries. Both incorporate pronunciation alternatives chosen by Broadcast News acoustic models. They differ in that the pronunciations in the first dictionary are selected from alternatives presented by CASS decision tree pronunciation models, while the second dictionary is selected from alternatives from a Broadcast News pronunciation model. The relative frequencies used in each dictionary can be reestimated through forced alignment over the training data. The training procedure also generates a *Soft reclustering of the HMM states* based on confusability found under the pronunciation model.

4.2. Performance of MBN Pronunciation Models

Clean, unaccented utterances (F0 condition) from the 1997 and 1998 HUB-4NE evaluation sets [14] were selected as the test set. The set contained 1263 utterances, with about 12,000 words. There were slightly more females than males, owing to the news anchors.

The pronunciation model to be evaluated is the *Reweighted and Augmented CASS Dictionary*. This dictionary consists of pronunciation alternatives for words, along with the number of observations of each alternative found in the 10 hour MBN training corpus. Two parameters control the amount of pronunciation variability introduced: a Minimum Count parameter discards variants that occur with less than a fixed number of occurrences; and a Minimum Relative Frequency threshold discards relatively infrequent alternates, regardless of how frequently or infrequently they are observed. Pronunciation probabilities are scaled so that the maximum for each word is 1.0. The performance of this dictionary with respect to these thresholds is given in Table 1. The second dictionary, estimated by finding the most frequent word pronunciations under the pronunciations produced by the Broadcast News decision tree pronunciation model, was also evaluated. However, it gave nearly identical results as the CASS-derived dictionary.

The second set of pronunciation models evaluated were HMMs whose mixture weights and means were reestimated after soft reclustering of the HMM states. Results are reported in Table 2. A significant 0.9% improvement in CER is found using the baseline dictionary. Furthermore, use of the reweighted and augmented dictionaries yields additional gains when used with these models for a 1.2% CER reduction.

5. CONCLUSION

The CASS Corpus of phonetically transcribed spontaneous Mandarin speech was used in a bootstrapping procedure to find pronunciation models for the Mandarin Broadcast News domain. Augmented dictionaries were used with soft reclustering of HMM states to reduce character transcription error rate. Direct acoustic measurements are incorporated as adjunct features in the decision tree pronunciation models, and the effectiveness of the predictive approach to pronunciation modeling was demonstrated in Chinese by modified the phonetic approach developed in English to the monosyllabic lexicons used in Mandarin. The improvements found are significant, but greater gains are expected in domains richer in spontaneous pronunciation effects, such as conversational speech. Additional documentation, data, and utilities can be found at www.cisp.jhu.edu/ws2000.

Acknowledgement We thank Michael Riley of ATT for use of the ATT large vocabulary decoder and pronunciation modeling tools; Murat Saraclar for helpful discussions; and Jun Wu for assistance in preparation of the language model training text. Mandarin Broadcast News data was obtained from LDC.

6. REFERENCES

- [1] LI Aijun, *et al.* CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In *Proc. ICSLP*, 2000.
- [2] M. Riley and A. Ljolje. Automatic generation of detailed pronunciation lexicons. In *Automatic Speech and Speaker Recognition: Advanced Topics*, pp. 285–302. Kluwer Academic Press, 1995.
- [3] W. Byrne, *et al.* Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In *Proc. ICASSP*, pp 313–316, 1998.
- [4] M. Riley, *et al.* Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29(2-4):209–224, November 1999.
- [5] Chen X., *et al.* An Application of SAMPA-C for Standard Chinese”. *Proc. ICSLP* 2000.
- [6] M. Saraclar. Pronunciation Modelling, Ph. D. Thesis, The Johns Hopkins University.
- [7] TIMIT acoustic-phonetic continuous speech corpus. Available from <http://www.ldc.upenn.edu/>.
- [8] S. Greenberg. Speaking in shorthand – a syllable centric perspective for understanding pronunciation variation. In *Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkraade, Netherlands, 1998.
- [9] S. Young, *et al.* The HTK book. Entropic Cambridge Research Laboratory, 1995.
- [10] SRILM - The SRI Language Modeling Toolkit. <http://www.speech.sri.com/projects/srilm/>.
- [11] A. Martin, *et al.* The DET curve in assessment of detection task performance. *EUROSPEECH*, pp. 1895-1898, 1997.
- [12] D. Talkin. A Robust Algorithm for Pitch Tracking (RAPT), Chapter 15, *Speech Coding and Synthesis*, Elsevier, 1995.
- [13] X. Luo and F. Jelinek. Probabilistic Classification of HMM States for Large Vocabulary Continuous Speech Recognition In *Proc. ICASSP*, pp. 2044-2047, 1999.
- [14] 1997 and 1998 HUB-4NE Mandarin Evaluations. www.itl.nist.gov/iaui/894.01/tests/ctr