INTERSPEECH 2009 BRIGHTON

Tonal Articulatory Feature for Mandarin and its Application to Conversational LVCSR

Qingqing Zhang, Jielin Pan and Yonghong Yan

ThinkIT Speech Laboratory
Institute of Acoustics Chinese Academy of Sciences
Beijing 100190, China
zhangqingqing, panjielin, yanyonghong@hccl.ioa.ac.cn

10.21437/Interspeech.2009-761

Abstract

This paper presents our recent work on the development of a tonal Articulatory Feature (AF) for Mandarin and its application to conversational LVCSR. Motivated by the theory of Mandarin phonology, eight features for classifying the acoustic units and one feature for classifying the tone are investigated and constructed in the paper, and the AF-based tandem approach is used to improve speech recognition performances. With this Mandarin AF set, a significant relative reduction on Character Error Rate is obtained over the baseline system using the standard acoustic feature, and the comparison between the ASR systems based on AF classifiers with and without the tonal feature demonstrates that the system with the tonal feature achieves better performances further.
Index Terms: Articulatory feature, tone, Mandarin, MLP

1. Introduction

Articulatory features, which are used to track the asynchronous alignment of the articulators which leads to the coarticulation phenomenon [1], have proven to be effective when used to improve the speech recognition performances. In recent years, more and more research activities, focus on how to make use of the AF in the ASR, are conducted. In the 2006 JHU Summer Workshop, an AF set which is based on articulatory phonology [2] adapted for pronunciation modeling [3] has been constructed. This AF set is tested on an English small-vocabulary set named SVitchboard [4], and the result shows that systems with the AF-based tandem feature achieve better performances than the baseline [5]. More recently, many researchers also report good results with JHU AF set [6] [7] [10] .

Every language has its speciality. This makes AF, which use phonetic properties to classify sounds, language dependent to a certain extent. For example, Mandarin is a tonal language, which benefits from explicitly modeling tones to distinguish ambiguous words [8]. If an AF set is constructed for describing each sound in Mandarin, tone has a crucial role. This is quite different from English and some other Western languages. Since the JHU AF set does not include the tonal feature, it is not especially appropriate for the tonal languages, such as Mandarin. Therefore, some language-dependent features need to be incorporated into the AF set for distinguishing the sounds of the corresponding language completely.

In this paper, we focus on improving Mandarin speech recognition accuracy using a tonal AF set for Mandarin in the spontaneous, conversational LVCSR. A Mandarin AF set, which includes eight features for the acoustic unit classification and one feature for the tone classification, is constructed based on the theory of Mandarin phonology, and the posteriors of these nine features derived from a multilayer perceptron (MLP) are used as features for the acoustic modeling, which is know as tandem approach [5]. In the experiment on the standard Mandarin CTS test set, compared with the baseline ASR using the standard acoustic feature, the tonal AF-based ASR has a 6.8% decrease on Character Error Rate, which performs better than the system without tonal feature. When the tonal AF tandem features are combined with the standard acoustic feature at the feature level and the word level, these combinations achieve better performances, which show the complementary information between these two features.

The rest of the paper is organized as follows: In section 2 the AF set for Mandarin is presented. In section 3, we describe the Mandarin conversational LVCSR system used in our experiment, and on the standard Mandarin CTS test set the performances between the baseline and Mandarin AF-based tandem features are compared and analyzed in section 4. Section 5 gives a brief conclusion of this paper.

2. Articulatory feature for Mandarin

Articulatory features (AFs) are abstract classes which characterize the most essential aspects of articulatory properties of speech sounds in a quantized form, leading to an intermediate representation between the signal and the lexical units [9]. Towards Mandarin, the AF set as

a basis for discrimination between words of Mandarin, should cover the most essential aspects of articulatory properties of Mandarin speech production. Even though the AF set such as JHU AF set can represent many properties of speech production in Mandarin, it is still incomplete. For example, when a sound belongs to "stop" or "fricative" in Mandarin, there is another distinctive feature named "aspiration" that we should consider to distinguish it from other sounds, which is different from English. Take "b" and "p" as an example, "b" and "p" are both stop sounds in English and Mandarin. The difference between these two sounds in Mandarin would be determined by calculating the probability that the first sound being UNASPIRATED and the second sound being ASPIRATED, while in English the difference would be that the first sound being VOICED and the second sound being UNVOICED [10].

Motivated by the analysis of articulatory phonology of Mandarin [11], a set of discrete multi-level AFs that can be used for classifying the sounds of Mandarin is investigated and constructed in our paper. Table 1 gives the feature specification, along with the values and the number of each feature.

Table 1: Specification of the Mandarin articulatory features. Note that each group also has 'silence' and 'reject' classes.

Feature name	Feature values	#
Place	Bilabial, labiodental, dental, alveolar, retroflex, alveolo-palatal, velar	9
Degree	Stop, nasal, fricative, lateral, affricate	7
Aspiration	Aspirated, unaspirated, other consonants	5
Glottal state	Voiced, unvoiced	3
Height	High, mid-high, mid, mid-low, low	7
Frontness	Front, mid-front, mid, mid-back, back	7
Rounding	Rounded, unrounded	4
Vowel	a, aa, ak, at, au, e, ea, ee, er, err, i, ii, ix, iy, o, u, uu, v,iaa, ioo, iee, iii, iuu, ivv	26
Tone	High-level, high-rising, low-dipping, high-falling	6

In each case of Table 1, the names of the features are given first, and then the values belonging to the feature are shown, followed by the numbers of the values. The first eight features are used for classifying the different consonants and vowels in Mandarin, and the last feature is to classify the tones marked on the acoustic units (Some researches consider Mandarin as a five-tone language including the neutral tone [8].). The meaning of each term used in the table refers to [12].

When the AF set is created, the feature transcriptions for training MLP can be generated by convert time-aligned phone-based labels to feature labels based on this AF set. This can be done using a canonically defined phone-feature mapping table [5]. A part of the mapping table, which is used to generate feature targets from time-aligned phone labels, is shown as an example in Table 2.

Table 2: Mapping from Mandarin phones "b", "p", "m" and "f" to their feature values in the Mandarin AF set.

Phone	b	p	m	f
Place	bilabial	bilabial	bilabial	labiodental
Degree	stop	stop	nasal	fricative
Aspiration	unaspirated	aspirated	other	other
Glottal state	unvoiced	unvoiced	voiced	unvoiced
Height	reject	reject	reject	reject
Frontness	reject	reject	reject	reject
Rounding	reject	reject	reject	reject
Vowel	reject	reject	reject	reject
Tone	reject	reject	reject	reject

3. Mandarin conversational LVCSR

In this section, we briefly describe the resources and data that are used for the development of the Mandarin conversational LVCSR system.

3.1. Acoustic Training and Test Data

The speech data used for training AF tandem features and all the acoustic models in the paper consists of about 120 hours of speech data from LDC database including CALL-FRIEND (20 hours), CALL-HOME (15 hours) and LDC04 training sets (90 hours) with a total of 1033 conversations (2066 sides) [13]. The test data [1] – HDev04 was collected by Hong Kong University of Science and Technology (HKUST) and released in 2005. It comprises of 4 hours of data with 24 phone calls. We separated it into two parts: HDev04-I and HDev04-II. HDev04-I with two and a half hours data is used as a holdout set to optimize the LM perplexity. HDev04-II with the residual one and a half hour data is used as our test set [18].

3.2. Dictionary and Language Models

The Mandarin dictionary used in our experiment consists of approximately 43,500 words and associated phonetic transcriptions. The phone set consists of 49 tone-

[1] The test set is not the same as the set named ldc04 in [14]: The former was released from LDC in 2005 with the Catalog No. LDC2005S15 [13], and the latter was released in 2004 as the development set in RT-04 NIST evaluation.

less phones (including SP and SIL) and associated tone markers[15].

The SRILM tools [16] are used to build the Trigram language models. The LM training data consists of four parts: general web data provided by the University of Washington [17], the transcription of the added CTS training data (swm03&ldc04) [18], self-collected data using Google with N-gram queries from the transcription of swm03 and ldc04 (search on swm03&ldc04), and self-collected general news corpora (news05) [18]. The HDev04-I is used as a holdout set to optimize LM perplexity. Table 3 shows the interpolation weights for the trigram LM.

Table 3: Interpolation weights for the trigram LM.

Source	Size(MB)	Weight
Washington	74	0.09
swm03&ldc04	3.90	0.55
search on swm03&ldc04	113	0.24
news05	1250	0.12

4. Experiments and Results

4.1. Baseline acoustic model

The baseline acoustic model is trained with the standard feature of Mel-Frequency Perceptual Linear Prediction (MFPLP) . This uses a reduced bandwidth analysis, 60-3400 Hz, to generate 12 MFPLP Cepstra along with the zeroth Cepstra. First and second-order differences were appended to give 39 features. Cepstral mean and variance normalization (CMN/CVN) was also applied per conversation side. The acoustic models used in the baseline system and throughout the paper are all state-clustered, crossword tri-phone HMMs with 12-component Gaussian mixture output densities per state. Table 4 shows the performance of the basic acoustic model with the standard MFPLP feature. This yields the Character Error Rate (CER) of 56.3% on the test set.

Table 4: 5xRT CER HDev04-II performance using MFPLP baseline and the Mandarin AF-based tandem features with and without the tonal feature.

Feature	CER (%)
MFPLP39	56.3
AF_NoTone-HLDA39	53.3
AF_Tone-HLDA39	52.5

4.2. Mandarin AF-based acoustic model

Multi-layer perceptrons (MLPs) have been successfully used for AF classification [5]. Here we use the Quicknet tools developed at ICSI [19] to train MLP, following standard procedures including taking a context window of 9 frames of MFPLP as input to a 3-layer MLP, and using cross validation data to determine the MLP learn rate and convergence. The number of MLP hidden units and the frame-level classification rates of cross-validation (CV) for the MLPs are reported in Table 5. The CV accuracy is measured against the forced-aligned labels, on a 10% subset of the training data that were set aside during MLP training.

Table 5: The number of hidden units / output units, and CV accuracy (%) for AF MLPs.

MLP Classifier	# of units	Accuracy(%)
Place	1900/9	76.51
Degree	1600/7	77.76
Aspiration	1400/5	78.68
Glottal state	1400/3	86.74
Height	1800/7	69.67
Frontness	1700/7	68.64
Rounding	1200/4	84.88
Vowel	2400/26	65.78
Tone	1700/6	63.61

As mentioned above, the tonal feature is quite important in the sense of speech discrimination in Mandarin. In order to analyze the impact of the tonal feature in the Mandarin speech recognition performance, two acoustic models are trained based on different AF sets: the AF set with the tonal feature, and the AF set without the tonal feature. As shown in Table 1, these two models have 74 and 68 feature dimensions respectively.

Table 4 gives the performances of these two acoustic models. "AF_Tone-HLDA39" and "AF_NoTone-HLDA39" refer to acoustic models based on AF classifiers with and without the tonal feature respectively. Both of these tandem features are subjected to apply HLDA to reduce the dimensions to 39, which are the same as the dimension of the baseline MFPLP. As can be seen, both of these two Mandarin AF-based acoustic models outperform the baseline significantly. In particular, AF_Tone-HLDA39 yields another 1.5% relative CER reduction compared to AF_NoTone-HLDA39, which proves that the tonal feature offers additional information beyond using the articulatory features for distinguishing the acoustic units directly.

4.3. Feature combination

Since the standard acoustic features and AF-based tandem features are quite different, the combination of both features might be beneficial as one feature might compensate for the errors made by the other system and vice versa [20]. Speech recognizers may be combined at var-

ious levels in the recognition process. Here, we concentrate on feature-level and word-level combinations, and Table 6 shows the performances of these combinations on the test set.

Table 6: 5xRT CER HDev04-II performance using the combinations of MFPLP and tonal AF-based tandem features on the feature-level and word-level.

Feature	CER (%)
MFPLP39	56.3
AF_Tone-HLDA39	52.5
MFPLP39+AF_Tone-HLDA39	51.5
3systems ROVER	50.6

In this table, "MFPLP39+AF_Tone-HLDA39" refers to the feature-level combination, which means concatenating the tonal AFs with MFPLP features at the input to the HMM system. After this combination, the concatenated vector has 78 features, whereas the stand-alone vector has 39 features. For the feature-level combination, there is a reduction on CER when compared to both of the two stand-alone features. "3systems ROVER" refers to the word-level combination using ROVER [21] on all of the three systems shown in the table above, and the word-level combination gives an additional 0.9% gain over the feature-level combination. These combinations consistently lead to reductions on CER, which proves that the Mandarin tonal AFs have significant complementary information with the standard acoustic feature MFPLP in the Mandarin speech recognition.

5. Conclusions

In this paper, a Mandarin articulatory feature set is developed for representing the properties of speech production of Mandarin. Based on this AF set, the speech recognition accuracy in the Mandarin conversational LVCSR system are improved significantly over the baseline with the standard acoustic feature MFPLP. As Mandarin is a tonal language, the impact of the tonal feature in the AF set is investigated as well, and the experiment results prove that the tonal feature incorporated into the AF set does help to improve the ASR performances.

6. Acknowledgements

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102), National Science & Technology Pillar Program (2008BAI50B00), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10874203, 60875014, 60535030).

7. References

[1] Erier, K., Freeman, G., "Using articulatory features for speech recognition", Communications, Computers, and Signal Processing, 1995. Proceedings. IEEE Pacific Rim Conference on 17-19 May 1995 Page(s):562 - 566

[2] C. P. Browman and L. Goldstein., "Articulatory phonology: An overview.", Phonetica, 49:155-180, 1992.

[3] K. Livescu., "Feature-based Pronunciation Modeling for Automatic Speech Recognition.", PhD dissertation, MIT, Cambridge, MA, 2005.

[4] S. King, J. Bilmes, and C. Bartels., "SVitchboard: Small-vocabulary tasks from Switchboard.", In Proc. Interspeech, 2005.

[5] K. Livescu et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition: JHU Summer Workshop Final Report," Technical report, Johns Hopkins University Center for Language and Speech Processing, 2007

[6] O. Cetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone MLPs," in Proc. ASRU, 2007, pp.36-41.

[7] O. Cetin et al., "An articulatory feature-based tandem approach and factored tandem observation modeling," in ICASSP, 2007

[8] X. Lei, M.-Y. Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR", In INTERSPEECH-2005, 2981-2984.

[9] Kirchhoff, K., "Robust speech recognition using articulatory information.", Ph.D. thesis, University of Bielefield.Kirchhoff, 1999

[10] J. Frankel et al., "Articulatory feature classifiers trained on 2000 hours of telephone speech", in Proc. INTERSPEECH, Antwerp, Belgium, August 2007.

[11] D. Zhou, Z. Wu, "PuTongHuaFaYinTuPu (in chinese)", The commercial press, Beijing, 1963

[12] Peter Ladefoged, "A Course In Phonetics", Third Edition, P11-P13 University of California, Los Angeles, 1993

[13] The Linguistic Data Consortium, http://www.ldc.upenn.edu./, 2009

[14] X. Liu, K.C. Sim, P.C. Woodland, M.J.F. Gales, B. Jia and K. Yu, "Development of the CUHTK 2004 RT04F mandarin conversational telephone speech transcription system", Proc. ICASSP'05, 2005.

[15] Q. Zhang, J. Pan and Y. Yan, "Mandarin-English bilingual speech recognition for real world music retrieval.", ICASSP 2008, paper 1147, Las Vegas, March 30-April 4, 2008

[16] SRILM - The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/, 2009

[17] T. Ng, M. Ostendorf, M. Y. Hwang, M. H. Siu, L. Bulyko and X. Lei, "Web-data augmented language models for mandarin conversatinal speech recognition", Proc. ICASSP'05, vol. 1, pp. 589-592, 2005.

[18] Jian Shao, Ta Li, Qingqing Zhang, Qingwei Zhao and Yonghong Yan, "A robust real-time decoder using memory-efficient state network," Transactions of IEICE on Information and Systems, Japan, 2008, E91-D(3): 529-537.

[19] David Johnson, http://www.icsi.berkeley.edu/Speech/qn.html, Speech Group at the International Computer Science Institute, berkeley, 2009

[20] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," Speech Communication, vol. 37, pp. 303-319, 2000.

[21] J. Fiscus. "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara, 1997.