FAST ACCENT IDENTIFICATION AND ACCENTED SPEECH RECOGNITION

LIU Wai Kat and Pascale FUNG

Human Language Technology Center
Department of Electrical and Electronic Engineering
University of Science and Technology
Clear Water Bay, Hong Kong
{eekat,pascale}@ee.ust.hk

ABSTRACT

The performance of speech recognition systems degrades when speaker accent is different from that in the training set. Accent-independent or accent-dependent recognition both require collection of more training data. In this paper, we propose a faster accent classification approach using phoneme-class models. We also present our findings in acoustic features sensitive to a Cantonese accent, and possibly other Asian language accents. In addition, we show how we can rapidly transform a native accent pronunciation dictionary to that for accented speech by simply using knowledge of the native language of the foreign speaker. The use of this accent-adapted dictionary reduces recognition error rate by 13.5%, similar to the results obtained from a longer, data-driven process.

1. INTRODUCTION

Most state-of-the-art speech recognition systems fail to perform well when the speaker has a regional accent different from that of the standard language the systems were trained on. Performance deteriorates further when the standard language is not the first language of the speaker. In Hong Kong, most people speak a particular version of Canto-English where their Cantonese is peppered with English words and their English has a particular local Cantonese accent. In [1], we point out several possible solutions for accent independent speech recognition. One is to train the system on a collection of speech database encompassing various accents. Another solution is to train accent-dependent recognizers using collected data. However, data collection in both cases is tedious and time-consuming. In addition, accent identification is needed for accent-dependent recognition.

A speaker is said to have an accent when s/he does not sound like a native speaker. Accent usually comes from the articulation habits of the speaker in her/his own native language. In learning a second (or more) language, the speaker has to learn a modification in the patterns of intonation, lexical stress, rhythm, grammar, as well as the use of additional distinctive phonemes [4]. Such modification leads to both acoustic and articulation differences. In this paper, we explore these aspects for (1) accent identification of and (2) accent-adaptive recognition of Hong Kong English.

We use the TIMIT corpus and a small HKTIMIT corpus for studying accent differences and for training. HKTIMIT is collected in our center and consists of 800 utterances from both native English speakers and Cantonese English speakers from our campus.

In this paper, we show how to perform fast accent classification using phoneme-class models instead of phoneme models, based on accent-sensitive features we discover. We also show a fast accent-adaptive method based on the knowledge of foreign speaker's own native language.

2. FAST ACCENT CLASSIFICATION BASED ON ACOUSTIC FEATURES

Most accent classification methods are based on accent-dependent models using common feature set[6] or feature-based discrimination[5, 3].

We propose a hybrid of using both feature-based and model-based discrimination. For fast accent classification using small amount of data, we do not use phoneme-based HMM for recognition. Instead, we train phoneme-class HMMs. The phoneme set is divided into six classes: 1) stops, 2) affricates, 3) fricatives, 4) nasals, 5) semivowels & glides and 6) vowels.

We investigate the following features and their first and second derivatives, for their effects on accent: fundamental frequency(F0), energy in rms value(E0), first formant frequency(F1), second formant frequency(F2),

third formant frequency(F3), and bandwidths of F1, F2 and F3, B1, B2 and B3 respectively.

The continuous speech is sampled at 16 kHz, high-frequency pre-emphasis is performed, Hamming windowed, followed by prosodic feature extraction on a frame by frame basis. Classification was based on a sequence of 3-state hidden Markov Models (HMM's) having single Gaussian densities.

The baseline system is built using all the 24 prosodic features. The baseline performance is 85.49% and 82.5% for close and open test respectively. By masking one feature at a time, we investigate its effect on accent classification on the training set. A best feature combination is used for the classifier.

The result shows the features in order of importance to accent classification to be: dd(E), d(E), E, d(F3), dd(F3), F3, B3, d(F0), F0, dd(F0), where E is energy, F3 is third formant, B3 is bandwidth of third formant, d() is first derivatives and dd() is the second derivatives. We explain the findings in the following sections.

2.1. Energy

Energy is an important feature that can show the differences of speaking style and structure of two different languages. Figure 1 gives the average mean energy of the phone classes between the two accent groups. The mean energy is higher for native English speakers in all classes. The variance is also much higher for native English speakers in all classes except affricates. This suggests the energy range for native speaker is higher. Figure 2 shows how the energy feature affects the performance.

Figure 1: Average mean energy of various phone classes

Phone classes	American	Cantonese
vowels	1035.1	506.65
nasals	601.70	252.28
stops	147.74	55.63
affricates	370.66	89.87
fricatives	224.79	117.99
semi-vowels & glides	1282.49	522.93

2.2. Formants

The second important parameter is the third formant together with its derivatives. Arslan and Hansen [5] suggest that F2 and F3 are both sensitive to accents, since their positions are shifted according to tongue movements. Tongue movements are supposed to be the most salient difference between native and non-native speakers. However, [5] shows that using F2 and F3 to classify accents work well for European accents but not for Asian accents. In our experiments, we find that only the formant position and bandwidth of F3 are important for classification between native and Hong Kong English accents, not those of F2. Figure 3 shows how formant features affect classification results.

Figure 2: accent classification accuracy with/without energy feature

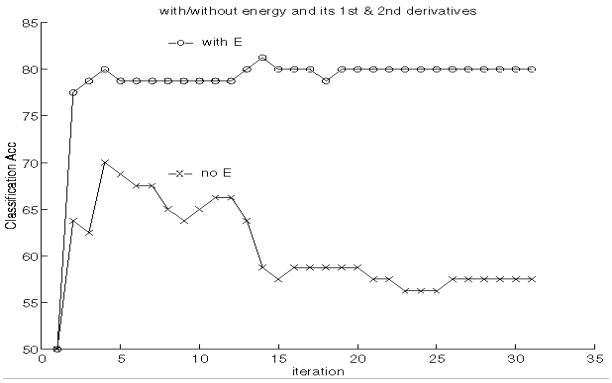


Figure 3: Classification error caused by formant features

features	error rate	features	error rate
full set	14.52%	full set	14.52%
no F1	14.14%	no B1	14.51%
no d(F1)	14.34%	no d(B1)	14.51%
no dd(F1)	14.13%	no dd(B1)	14.36%
no F2	14.21%	no B2	14.08%
no d(F2)	14.37%	no d(B2)	14.39%
no dd(F2)	14.22%	no dd(B2)	14.33%
no F3	15.13%	no B3	14.72%
no d(F3)	15.14%	no d(B3)	14.33%
no dd(F3)	15.14%	no dd(B3)	14.46%

2.3. Fundamental frequency

Human perception tests indicate that the listeners based their accent classification decisions partly on prosodic features such as pitch movements, rhythm and pausing [3]. We find that the pitch contours of Cantonese speakers are choppier (Figure 4).This result can also be reflected by the fact that average number of countable voiced region is greater and the average duration per

voiced region is smaller for Cantonese speakers. Cantonese language is a monosyllabic. Syllables in Cantonese are made up of an Initial and a Final. Figure 4 shows that speakers has carried their first language speaking style to foreign language. In our feature experiment, we find that if ignoring F0, its first derivatives and its second derivatives are masked, there is an increase of 5.6% in accent classification error rate.

Figure 5: F0 information reduces classification error

parameter set	error rate
full set	14.52%
no dd(F0)	14.58%
no F0	14.65%
no d(F0)	14.67%
no F0 info.	15.33 %

3. FAST ACCENTED SPEECH RECOGNITION BASED ON NATIVE LANGUAGE KNOWLEDGE

The above analysis on prosody information only show the acoustics differences between different accent groups. These features, while powerful for accent classification, are difficult to incorporate into accent adaptation. We turn to another major difference between native and non-native speakers–pronunciation difference. [2] shows that it is effective to incorporate accent-specific pronunciation rules into the dictionary for recognition. A phoneme A in the speech of a native speaker can be mapped to the phoneme B in the speech of a non-native speaker.

The information of such mapping rules can be obtained by three sources:

The first source is from the position of phonemes in F1-F2 plan. Figure 6 shows mean F1 vs F2 frequencies of the vowels for native American English and Cantonese accented English. This method can show the degree of differences of the phoneme between two accent classes and in what direction a phoneme is moving towards another one. For example, from Figure 6, the UW sound for Cantonese speakers is far away from that of native American speaker and it is moving towards the sound OW. The AA sound of both groups are more overlapping . The same result can also be found in the phoneme recognizer output. However, this method cannot show phoneme deletion and phoneme insertion but only phoneme substitution. It can be applied to transformation-based accent-adaptation methods.

Figure 6: A first formant vs second formant plot for vowels

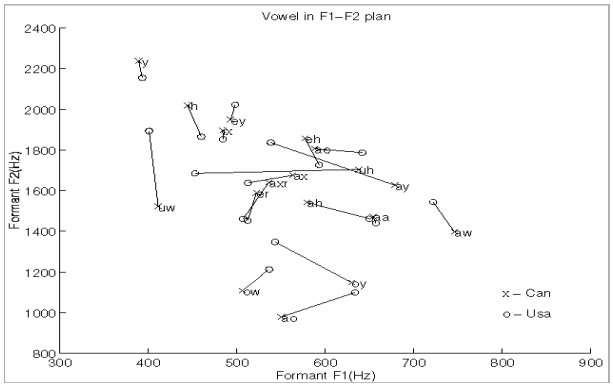


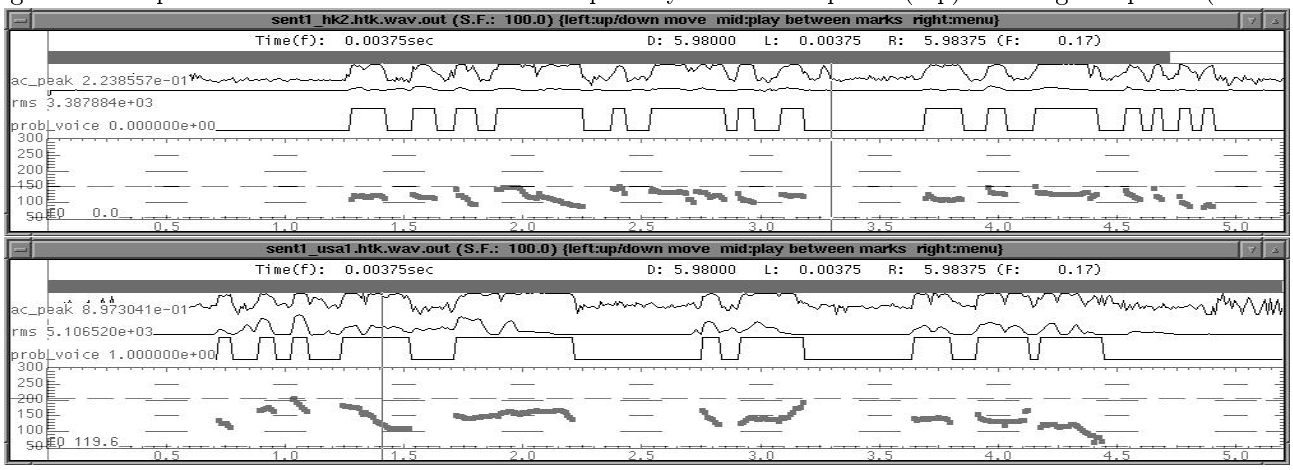
Figure 7: Word accuracy of using native English dictionary and dictionary adapted to Cantonese

Speaker	native dict	accented dict
spk01	74.07%	75.93%
spk02	64.81%	69.44%
spk03	70.37%	76.85%
spk04	67.59%	71.30%
average	69.21%	73.38%

The second method is data-driven. Non-native accent speech is passed to a native accent phoneme recognizer and the result can be a confusion matrix showing the general mapping of phonemes between the two sets. The most confusable sounds in Hong Kong English are /R/, /AXR/, /P/, /ER/, /K/ and /G/. Those sound are either missing or seldom occur in the Cantonese language.

We propose a new method to derive mapping rules using linguists knowledge. This method is the fastest and easiest for making the accent-specific dictionary. Moreover, such kind of knowledge is well studied and less data-dependent. For example, linguists have shown that there are some sounds that do not occur in Cantonese such as /AXR/, /AX/, /AE/, /IH/, /AH/ and /UH/. In our system, we apply 28 phonetic rules to an electronic dictionary (BEEP) designed for native English speakers. The dictionary size is doubled. In Figure 7, we can see that accented speech recognition results are better by using the accent-adapted dictionary than using the native pronunciation dictionary, giving an average of 13.5% error rate reduction.

Figure 4: The pitch contour of the same utterance spoke by Cantonese speaker(top) and English speaker(bottom)



4. CONCLUSION

In this paper, we demonstrate which acoustic features are important for accent classification of Hong Kong English. We show that in general, energy, formant and fundamental frequency information are the most discriminative features for identifying a Cantonese (and possibly other Asian) accents. We also show that, unlike for European accents, only F3, instead of both F2 and F3 [5], is indicative of a Cantonese (and possibly other Asian) accent.

We also show the recognition results of accented speech by using a knowledge-based accent-specific pronunciation dictionary. We obtain this knowledge from exploring the native language characteristics of the foreign accent speaker. We show that we can reduce the error rate from around 30.89% to 26.62%, similar to the reduction from 30.9% to 24.8% reported in [2]. Our method is much faster than those obtained from database as in [2].

5. REFERENCES

[1] Pascale Fung, Bertram Shi, Dekai Wu, Lam Wai Bun, and Wong Shuen Kong. Dealing with multilinguality in a spoken language query translator. In *Proceedings of ACL 97 Workshop on Spoken Language Translation*, pages 40–43, Madrid, Spain, July 1997.

[2] Humphries J.J. and Woodland P.C. Using accent-specific pronunciation modeling for improved large vocabulary continuous speech recognition. In *Proc. ICASSP'97*, 1997.

[3] Kumpf K. and King R. W. Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In *Proc. EUROSPEECH '97*, pages 2323–2326, 1997.

[4] Hansen J. H. L. and Arslan. L. M. Foreign accent classification using source generator based prosodic features. In *Proc. ICASSP '95*, pages 836–839, 1995.

[5] Arslan L. M. and Hansen H. L. Frequency characteristics of foreign accented speech. In *Proc. ICASSP'97*, pages 1123–1126, 1997.

[6] Trancoso I. Teixeira C. and Serralheiro A. Accent identication. In *Proc. ICSLP'96*, pages 1784–7, 1996.