MAM: A MULTI-ACCENT MANDARIN CORPUS

Chung-Hsiang Huang, Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan

m983040007@student.nsysu.edu.tw, cpchen@cse.nsysu.edu.tw

ABSTRACT

This paper describes the design, the collection, and the labeling of a Multi-Accent Mandarin corpus called MAM. MAM is constructed primarily for automatic Chinese accent detection. We aim to study the idiosyncrasies of the Chinese accents of people from different regions of the world. The text in the corpus is designed to be easily readable for international Chinese-speaking students studying in Taiwan, while maintaining sufficient linguistic complexity. The speech data in MAM is collected from overseas students currently enrolled in the National Sun Yat-sen Univeristy. There are 30 speakers with 40 utterances each, so a total number of $1,200$ utterances are collected. The recruited speakers are geographically divided into three groups: 10 speakers (6 male and 4 female) are from Indonesia, 10 speakers (5 and 5) are from Malaysia, and 10 speakers (5 and 5) are from Hong Kong or Macau. In addition, we have selected 10 speakers (5 and 5) from Taiwan in the TCC-300 corpus. The utterances in MAM are manually analyzed for actual pronunciation, and a rudimentary accent classifier based on the pronunciation variation patterns across different geographic areas is constructed.

Index Terms— multi-accent Mandarin corpus, accent detection

1. INTRODUCTION

Accent is the unique pronunciation manner depending on the geographical location due to the influence of different mother languages (first languages), which can be dialects or completely different languages. For example, people from Japan speaks Chinese with a special flavor, and similarly for people from the United States. The Mandarin spoken by people with Taiwanese as the mother tongue is often called Taiwan-Mandarin, similar to the case that the English spoken by people in Singapore is also known as the Singlish. Even in Taiwan, there is perceptible difference between the Mandarin spoken by Min-Nan and Hakka. The reasons that produce *pronunciation differences* may be numerous, so we want to find a solution of this problem.

Accent detection tasks have been carried out in the case of English. In [1], the American and Indian accented English are chosen and classified. Another work in [2], the researchers identified 3 major accents in Australia. A multi-accent English news speech corpus is collected from 6 regions in [3], which are US, Great Britain, Australia, North Africa, Middle East and India.

Mandarin is one of the mostly used languages in the world with more than one billion speakers. The booming of China means that more and more people speaks Mandarin in the world, so Mandarin is a global language. For spoken language technology applications, the accent of Mandarin becomes an important and practical research topic with many related research works. For instance, accented Mandarin speech recognition based on the method of gradient tree boosting is introduced in [4]. In [5], three different Chinese dialects are collected to build an accent detection system.

Although there are many public-domain Mandarin corpora collected in Taiwan for the advance of speech and language technology, e.g. [6, 7, 8] released by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), most of these corpora are mainly collected for the purposes of automatic speech recognition or speech synthesis. In order to study of accent via corpus, we need to create a new multi-accent corpus. Such a corpus provides instances of acoustic waveforms, and facilitate a fair comparison of different techniques for accent-related evaluation tasks. For the demand of accent classification, we design and construct a multi-accent Mandarin corpus, and collect speech data from the Mandarin-speaking international students from Malaysia, Hong Kong, Macau, and Indonesia in campus.

The rest of this paper is organized as follows. The design of MAM is introduced in Section 2. The collection of the MAM data is detailed in Section 3. We describe how the recorded utterances are analyzed, and summarize the corpus statistics in Section 4. The results of using MAM

978-1-4577-0931-9/11/$26.00 ©2011 IEEE

Table 1. The numbers of the international students of each region in NSYSU.

Malaysia	Hong Kong	Macau	Indonesia
50	16	40	11

in accent detection are presented in Section 5. Lastly, the conclusion of our current works is given in Secton 6.

2. CORPUS DESIGN

Our goal in MAM is to collect sufficient enough speakers and speech data from every chosen accent groups. Thus, the guidelines of corpus design are

- at least 10 speakers in each region, and 40 sentences for each speaker
- the text prompt should be easily readable to avoid unnecessary complications
- the linguistic complexity should be maintained at a certain level

The numbers of the international students in the National Sun Yat-sen Univerisity from various countries satisfying the constraint on the number of speakers per region are shown in Table 1.

The text source of speeches can be split into two parts. The first part is selected from the text prompts used in [9], and second part is harvested from on-line Chinese news websites and Chinese newspapers. The number of characters in each utterance ranges from 10 to 30. In addition, we keep the total number of characters in each speaker group balanced. We ensure that the text is free of ambiguity or confusion to the speakers.

For illustration, we excerpt two sentences below.

1. 那些錢是要用來幫助窮人的
2. 一個有靈活腦袋的孩子將會比一個愚蠢的小孩有出息

The first sentence means *The money is used to help the poor*. In Mandarin, 錢(*money*) and 窮人(*poor*) are common words, therefore they are easy to read. The second sentence means *A smart child will be more successful than a foolish child*. Here, the words 靈活(*smart*) and 愚蠢(*foolish*) are much harder than the previous ones, and 出息(*successful*) is not as commonly seen. Such words will make the sentence harder to read.

3. DATA COLLECTION

3.1. Speaker Groups

Based on Table 1, we decide to use three speaker groups: Indonesia, Malaysia, and Hong Kong-Macau. Since the main language in Hong Kong and Macau is Cantonese, we put them into the same group.

The background knowledge about the mother languages of these groups will be helpful to improve the integrity of our corpus. In Indonesia, there are more than 7 different languages/dialects. In Hong Kong and Macau, the most common language is Cantonese. Portuguese is used in Macau before, but few teenagers can speak it nowadays. Bahasa Malaysia and English are both official languages in Malaysia, and Bahasa Malaysia is very similar to Indonesian. When speaking Mandarin, the mother languages of the speaker will obviously affect the accent. This phenomenon will leads to non-standard Mandarin, and we can easily distinguish the difference by human ear.

It is important to make sure that the speakers are representative of the accent. If most of the pronunciation differences in speech happen from misreading rather than from accent, the speech data will not be suitable for accent detection. We make sure that the mother languages of the speakers for a given region are the same, and these mother languages differ from region to region. We also make the speaker composition (e.g. gender, age) to be as similar as possible across the groups.

3.2. Recording Setup

We use Cool Edit Pro 2.0[1] as recording software. The environmental noise is ensured to be lower than 40 dB signal-to-noise ratio, to avoid affecting the quality of speech. We use Audio-technica Stereo microphone AT9941 to record the speech. The discrete-time sampling rate is $16,000$ Hz.

Before the recording, we give the speaker the text prompt to check the readability.[2] Each utterance is recorded as one audio file. We check each utterance until it is successfully recorded. The average time for recording per speaker is approximately two hours. We also inquire the speakers of their mother languages and daily used languages, how long they have learned Mandarin, and the difficulty level they feel about reading the text. The whole speech recording flow is shown in Fig. 1. Note that the utterances of Mandarin from Taiwan speakers are selected TCC300 [7].

[1] Now this software is called ”Adobe Audition”, see `http://www.adobe.com/special/products/audition/syntrillium.html` for more information.

[2] We require a speaker to be able to read at least 80% of the text.

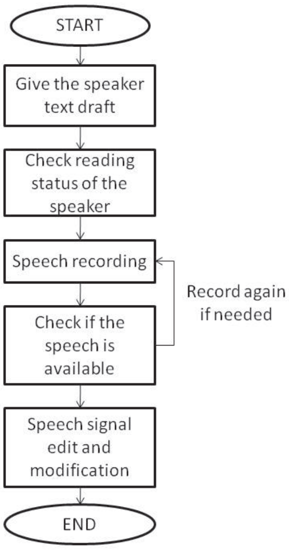
START
Give the speaker text draft
Check reading status of the speaker
Speech recording
Record again if needed
Check if the speech is available
Speech signal edit and modification
END

Fig. 1. The flow chart of our speech recording work.

4. CORPUS CONSTRUCTION

The recorded $1,200$ utterances from non-Taiwan speakers are manually transcribed. We also choose speech files of 10 speakers (5 male and 5 female) from the TCC-300 corpus [7], and balance the sum of characters of each group. We have $25,226$ spoken character tokens in MAM, consisting of $5,759$ in Hong Kong-Macau group, $6,107$ in Malaysia group, $7,006$ in Indonesia group, and $6,354$ in Taiwan group. The total size of these audio files is about 634 MB.

4.1. Post-Processing

The groups are abbreviated as: IN for Indonesia, HM for Hong Kong and Macau, MY for Malaysia, and TW for Taiwan. Although there still have several different accents in Taiwan as we discussed in Section 1 before, we use TW to represent the whole accents in Taiwan to simplify the goal. We use 80% (8 speakers, 320 utterances) of the data per group as the training data, and the remaining 20% (2 speakers, 80 sentences) as the test data.

Each utterance is further segmented. We segment an utterance into characters, and segment one character again into phones which are corresponding to the Mandarin phonetic symbols. The Mandarin phonetic symbol set consists of 37 symbols. Each character can be phoned by $1-3$ Mandarin phonetic symbols. Furthermore, since Mandarin is a tonal language, we also label the tones of each characters. There are 5 tones in Mandarin language, so we transcribe the tones with digits 1 to 5.

The main goal of using MAM is to learn the differences between accents. The differences are divided into several categories as follows.

Table 2. The count of pronunciation differences in MAM.

Region	A1	A2	A3	B1	B2	C	D
HM	102	11	222	40	223	83	16
MY	94	9	53	9	58	17	7
ID	94	4	43	19	422	48	21

- **Type A (Phone)** In our observation, many retroflex consonant in foreign accent speeches are pronounced like alveolars. For example, a character is pronounced as “zhang (張)”, however speaker may pronounce this character as “zang (髒)”. The phoneme /zh/ is pronounced as /z/. Such a phonetic difference is said to be of Type A. Moreover, Type A is further divided into three subtypes: A1 for a retroflex consonant is pronounced as non-retroflex, A2 for an non-retroflex (usually alveolars) consonant is pronounced as retroflex, and A3 for other situations, such as the difference happens in initials, finals, or medials. There still can be classified to more sub-groups in type A3 but it will be too exhaustive. For this reason, we only focus on retroflexes.

- **Type B (Tone)** A character dui4 (對) of the 4th tone can be pronounced as dui1. Such a phonetic difference is said to be of Type B. Type B is further divided into two subtypes: B1 if the pronounced tone exists for the character, and B2 for tone misreading or tone sandhi.

- **Type C** fast reading, repetition, filler, affected by front or back, or in the situation that differences happen in both phone and tone. All of these differences are of Type C.

- **Type D** misreading, unclear pronunciation, pause, or some situations that can not be classified. These situations above are are of Type D.

For example, the utterance 雖然他們的功能不同，但是運作的方式基本上是一樣的is segmented as “雖su(sui), A3”, “不bu1(bu4), B2”, “但是da1si1(dan4shi4), C, C”, “作zuo2(zuo4), B2”, “式si1(shi4), C”, and “是si(shi), A1”.

We list the pronunciation difference in MAM in Table 2. From Table 2, we can see that there are frequent Type A differences in HM group, and that the ID and MY groups have frequent Type B differences.

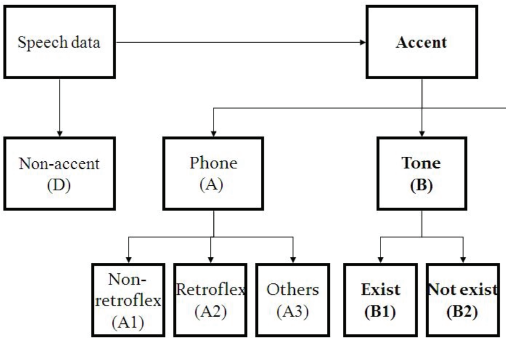

Fig. 2. Classification tree according to pronunciation differences in corpus.

5. CORRESPONDING EXPERIMENT

In this section, we will discuss that how we used the MAM corpus in our research. As the description in Sec. 1, we will use MAM to build experiments of accent classification. How to use the informations in hand, such as pronunciation characteristics to process accent classification and identification is our main goal. As in the researches of [1, 10], they use Gaussian Mixture Model (GMM) as the method of classification. GMM is a well-known method and is easy to use, however the accuracy is not high enough. To solve this problem in [11, 12], the researchers use Support Vector Machine to improve the accuracy and reduce error rate.

In our research, we discussed about the pronunciation differences in Sec. 4.1. We found out that there exist some distributions of pronunciation differences between foreign regions. Different accents have different pronunciation characteristics, therefore, we can classify an unknown accent by its pronunciation characteristics. Owing to the characteristics of pronunciation differences, it is possible to decide which accent is most likely to produce an utterance, if the type-dependent error probabilities for each accent are estimated.

Let K be the types of errors ($K = 7$). Let $\mathbf{n} = (n_1, \ldots, n_K)^T$ be an integer-valued vector where the i^{th} component indicates the number of errors in error type i. The likelihood that an utterance with error vector $\mathbf{n}$ being uttered by a speaker from accent a_j can be written as

$$P(\mathbf{n}|a_j) = \prod_{i=1}^{K} p(n_i|a_j) = \prod_{i=1}^{K} \frac{\lambda_{ij}^{n_i} e^{-\lambda_{ij}}}{n_i!}, \quad j = 1, 2, 3. \tag{1}$$

In (1), we make the assumption that the numbers of errors in different error types are independent and Poisson-distributed with parameter $\{\lambda_{ij}\}$. These parameters can be estimated from a training set. The empirical numbers of error counts on MAM are shown in Table 2.

Table 3. The pronunciation differences and corresponding λ_{ij}.

Region	HM	MY	ID
A1	0.228	0.225	0.213
A2	0.028	0.016	0.003
A3	0.425	0.128	0.106
B1	0.09	0.016	0.05
B2	0.413	0.125	1.288
C	0.15	0.04	0.138
D	0.044	0.009	0.047

In our experiment, the Taiwan accent was set to be the baseline as default. We chose 80% (4 male and 4 female per group, 320 sentences) of the speech data as training set, and 20% (1 and 1 per group, 80 sentences) of them as test set. We estimated the Poisson parameters from the training set. Table 3 shows the λ_{ij} of each pronunciation differences in the experiment. Then we calculated the probability (1) for each accent. The result of classification accuracy of 3 groups is 52.5% for HM, 32.5% for ID, and 33.75% for MY.

6. CONCLUSION

In this paper, we describe the design and whole construction processes of our multi-accent Mandarin corpus, the MAM corpus. Until now, 3.2 hours of data were collected and still extensible. Over 24,000 characters are used in the text, and about 10,000 characters are not repeatedly used. In our experiment, we use the MAM corpus in accent classification and speech segmentation. Poisson model is trained in each accent group in training set, and used to classify accent in test set. The distribution of test data will close to one of the group in train set, and we can recognize the accent of test data.

The construction of MAM corpus can still be improved. We know that the Mandarin learning levels of every speakers are not the same, so we should design our text draft before and after speech recording, to make sure that the reading problems will seldom happen.

Not only for accent classification, MAM can be also used in speech recognition or in accented speeches synthesis. We have already known that accent problem will critically impact the result of speech recognition. MAM can help to solve accent problems which may occur. Another possible research topic is text-to-speech system, according to the speech recognition technologies. The MAM corpus is a good linguistic resource for natural and foreign accented Mandarin.

7. REFERENCES

[1] Shamalee Deshpande, Sharat Chikkerur, and Venu Govindaraju, “Accent classification in speech,” in *Automatic Identification Advanced Technologies*. IEEE, 2005, pp. 803–806.

[2] P. Nguyen, D. Tran, X. Huang, and D. Sharma, “Australian accent-based speaker classification,” in *Proceedings of the 2010 Third International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 416–419.

[3] D. Vergyri, L. Lamel, and J.L. Gauvain, “Automatic speech recognition of multiple accented english data,” in *Proc. Interspeech*, 2010, pp. 1652–1655.

[4] Ming-Chin Yen, Po-San Lai, , and Jui-Feng Yeh, “Accented speech recognition based on gradient tree boosting with duration and articulation features,” in *Rocling*, 2010.

[5] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S. Y. Yoon, “Accent detection and speech recognition for shanghai-accented mandarin,” in *Proc. Interspeech*, 2005.

[6] Hsiao-Chuan Wang, “Mat–a project to collect mandarin speech data through telephone networks in taiwan,” *International Journal of Computational Linguistics and Chinese Language Processing*, pp. 73–89, 1997.

[7] ACLCLP, “Tcc-300edu,” introduction: `http://www.aclclp.org.tw/doc/tcc300_brief.pdf`, 2005.

[8] Academia Sinica, “Sinica corpus,” introduction: `http://db1x.sinica.edu.tw/kiwi/mkiwi/`, 1994.

[9] Chih-Yung Yang and Chia-Ping Chen, “A Hidden Markov Model-based approach for emotional speech synthesis,” in *7th ISCA Speech Synthesis Workshop 2010, Kyoto, Japan*, 2010.

[10] Tao Chen, Chao Huang, Eric Chang, and Jingchun Wang, “Automatic accent identification using gaussian mixture models,” in *IEEE Workshop on ASRU*, 2001, pp. 343–346.

[11] P. Watanaprakornkul, C. Eksombatchai, and P. Chien, “Accent classification,” in *Machine Learning Final Projects*, 2010.

[12] Jue Hou, Yi Liu, T.F. Zheng, J. Olsen, and Jilei Tian, “Multi-layered features with svm for chinese accent identification,” in *Audio Language and Image Processing (ICALIP), 2010 International Conference on*, 2010, pp. 25–30.