Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research

Eric Chang, Yu Shi, Jianlai Zhou, and Chao Huang

Microsoft Research China
Beijing, China
{echang, i-yshi, jlzhou, chaoh}@microsoft.com

Abstract

The necessity of gathering data has been an impediment for researchers and students who are interested in getting started in the fields related to speech recognition. We are proposing a new approach of distributing data that is designed to quickly help researchers and students achieve a set of baseline results to build upon. Furthermore, by leveraging publicly available programs, all researchers will be able to exactly reproduce results that are described in this paper. We also aim to facilitate comparison of recognition results in the field of Mandarin speech recognition by including a testing set in the toolbox. We describe a toolbox that includes Mandarin speech data from 125 speakers, suitable language model, scripts and data files required for recreating a set of baseline experiments, and a copy of Microsoft SAPI 5.0 SDK that can help professors and students who wish to jumpstart research programs in speech technologies. By lowering the barrier of entry to the field, we hope to encourage more participation in the study of Mandarin speech recognition.

1. Introduction

The availability of data is a prerequisite for conducting speech recognition research. However, while many students and professors would like to participate in speech recognition research, the task of collecting a sufficient amount of data for training and testing acoustic models is an obstacle. Recently, various speech recognition related tools and programs have become available [1][2][3]. In this paper, we describe our plan to release a toolbox that would enable individual professors or students to quickly begin research in Mandarin speech recognition, dialog systems, and multimodal user interface. The toolbox contains acoustic data, a syllable based language model, scripts and experimental results for baseline experiments, and a production quality speech recognition and synthesis software development kit. The toolbox is thus valuable to researchers in the area of speech recognition, user interface, and voice based interactive systems. There have been other Mandarin speech corpora organized in the past, such as the Mandarin Broadcast News database available from Linguistic Data Consortium [4], the Mandarin dictation corpora from the 863 project [5], HKU 96 and 99 corpora [6], and Mandarin Across Taiwan [7]. However, it is still relatively difficult for professors and students in China with limited budgets to get access to these databases. We hope that by providing this toolbox, more professors and students will participate in speech recognition research. This paper is organized as follows: In Section 2, we describe the contents of the toolbox. In Section 3, we provide the baseline experiments along with their implementations and results. Section 4 briefly describes the capabilities of SAPI 5.0 SDK in addition to its current uses. Lastly, Section 5 provides the conclusion of the paper.

2. Toolbox content description

The goal in designing this toolbox is to provide a suite of components that are useful to researchers in speech recognition, dialog systems, and user interfaces. Also, the size of the database is designed to be suitable for students who do not have very powerful computers with large hard disks. Thus, the database should also be useful for students who want to explore research in speech recognition.

2.1. Content of the toolbox

We plan to provide researchers and students access to the toolbox for research use. The content of the toolbox are listed below:

- Read speech from 100 male speakers, each speaking approximately 200 sentences, for a total of 19,688 sentences and 454,315 syllables.
- Manually verified syllable transcriptions for the waveforms.
- An additional test set of 25 male speakers, with 20 test sentences per speaker.
- Manually verified transcriptions for the test set.
- Bigram based tonal syllable language model.
- Data files required to train and test tonal syllable acoustic models, with and without the use of language models.
- A complete set of scripts which utilize the publicly available HTK toolkit to train and test Mandarin syllable acoustic models.
- Results and log files from baseline experiments.
- Microsoft SAPI 5.0 SDK, containing Mandarin Chinese, English, and Japanese speech recognition reference engines and Mandarin Chinese and English text to speech engines.

2.2. Waveform corpora description

All waveforms were recorded at a sampling rate of 16,000 Hertz and 16 bits per sample using close talking microphones connected to Creative Lab Soundblaster cards in quiet office environments. The recording software enforces a requirement on the loudness of the recorded speech so that no clipping

occurs. During recording, an experienced verifier sits beside the speaker to make sure that all scripts were properly pronounced.

2.2.1. Training set waveform corpora

The training set corpus was collected from students at a broadcasting school in Beijing, so most participating speakers were capable of enunciating clearly. We have compiled the speakers' place of origin (as provided by each subject) as shown in Table 1. The age of the subjects range from 18 to 40, with the vast majority under 25. The scripts spoken by the speakers were carefully designed to cover the Mandarin Chinese syllable set. Most of the sentences were selected from newspaper texts, while 30 sentences were created to be phonetically balanced and to cover the Mandarin syllable set.

2.2.2. Testing set waveform corpora

The testing set waveforms were recorded internally within Microsoft's Beijing office from volunteers who are considered to have the Beijing dialect. The scripts spoken were sentences taken from newspaper text that had perplexities in the range of 100 to 200.

Dialect	Number of Speaker
Anhui	2
Beijing	33
Chongqing	2
Fujian	2
Guangxi	3
Henan	4
Hebei	6
Heilongjiang	4
Hunan	2
Jiangsu	4
Jiangxi	1
Jinin	1
Liaoning	7
Nanjing	1
Neimonggu	3
Qingdao	1
Qinghai	3
Shanxi	7
Shandong	5
Shanghai	2
Sichuan	1
Tangshan	1
Wuhan	1
Xingjiang	1
Yunnan	2
Unknown	1

Table 1: Number of speakers per each dialect in the training set corpus.

3. Baseline experiments

In this toolbox, we have provided the scripts and the data files necessary to reproduce a set of baseline training and recognition experiments so that any student should be able to reproduce our results on a standard PC (the whole run from start to finish takes approximately 62 hours to complete on a 866 MHz. Pentium III machine with 512 MB of memory running Windows 2000 Server Edition).

3.1. Training

The whole training procedure closely follows the one outlined in the HTK manual. The feature used is a 39 order feature vector, consisting of 12 cepstral coefficients, energy, and their first and second order differences. The feature vector is calculated using a window size of 25 ms and a step size of 10 ms. The whole training procedure should be divided into two stages: monophone and triphone. In each stage, there are always two steps which are repeated iteratively: estimation and realignment. The process begins with the training of the monophone models, followed by training of the triphone models. For predicting unseen triphone in recognition, the parameter of tied-state triphone should be estimated. A detailed and specific case using Mandarin will be described as follows.

3.1.1. Monophone model training

To train the acoustic models, we use the syllable based approach. The basic acoustic units used for recognition are shown in Table 2 [8]. The baseline acoustic model was designed to be tonal since tone is an important feature of the Chinese language. Monophone models were first created using all 19,688 sentences. Since the transcription of the waveforms is a lexical transcription rather than a phonetic transcription, after the initial set of models have been trained, we performed force alignment of each waveform against its transcription, but allowing for the insertion and deletion of silences and short pauses between syllables. This improved transcription was used for all further training.

Initial	b, c, ch, d, f, g, ga, ge, ger, go, h, j, k, l, m, n, p, q, r, s, sh, t, w, x, y, z, zh
Tonal Final	a(1-5), ai(1-4), an(1-4), ang(1-5), ao(1-4), e(1-5), ei(1-4), en(1-5), eng(1-4), er(2-4), i(1-5), ia(1-4), ib(1-4), ian(1-5), iang(1-4), iao(1-4), ie(1-4), if(1-4), in(1-4), ing(1-4), iong(1-3), iu(1-5), o(1-5), ong(1-4), ou(1-5), u(1-5), ua(1-4), uai(1-4), uan(1-4), uang(1-4), ui(1-4), un(1-4), uo(1-5), v(1-4), van(1-4), ve(1-4), vn(1-4)

Table 2: Initial and tonal final units used for acoustic modeling. (numbers following final units indicates the range of tones represented).

3.1.2. Expansion to context dependent triphone models

After the monophone models are trained, all possible triphone expansions based on the full syllable dictionary are performed. In order to account for cross syllable contexts, the HTK configuration parameters ALLOWXWRDEXP and FORCECXTEXP were both set to TRUE. This results in a total of 295,180 triphones. Out of these triphones, 95,534 triphones actually occur in the training corpus. Each triphone model is represented by one single Gaussian.

After performing several iterations of embedded reestimation, we use the decision tree based clustering

capability of the HTK toolkit to tie similar states of triphones to each other. A subset of decision tree questions are listed in Table 3. These questions are designed to guide the clustering of the Gaussian mixtures based on the contexts of each phone model when little data is available. After clustering, the number of unique Gaussian mixtures is reduced to 2,392. We then use the HTK toolkit's Gaussian splitting capability to incrementally increase the number of Gaussians per mixture to 8 Gaussians per mixture.

```
QS "QS_0" {r-*,m-*,n-*,l-*,y-*,w-*}
QS "QS_1" {*+r,*+m,*+n,*+l,*+y,*+w}
QS "QS_2" {b-*,p-*,m-*,w-*}
QS "QS_3" {*+b,*+p,*+m,*+w}
QS "QS_4" {f-*}
QS "QS_5" {*+f}
QS "QS_6" {zh-*,ch-*,sh-*,r-*}
QS "QS_7" {*+zh,*+ch,*+sh,*+r}
QS "QS_8" {z-*,c-*,s-*}
QS "QS_9" {*+z,*+c,*+s}
QS "QS_10" {d-*,t-*,n-*,l-*}
QS "QS_11" {*+d,*+t,*+n,*+l}
QS "QS_12" {j-*,q-*,x-*,y-*}
QS "QS_13" {*+j,*+q,*+x,*+y}
QS "QS_14" {g-*,k-*,h-*}
QS "QS_15" {*+g,*+k,*+h}
QS "QS_16" {ga-*,ge-*,go-*,ger-*}
QS "QS_17" {*+ga,*+ge,*+go,*+ger}
QS "QS_18" {ga-*,ge-*,go-*,ger-*,y-*,w-*}
QS "QS_19" {*+ga,*+ge,*+go,*+ger,*+y,*+w}
QS "QS_20" {ga-*}
```

Table 3: Sample question files used for clustering Gaussian clusters when little data is available. The full list will be included in the toolbox.

3.2. Recognition experiments

After the acoustic models are trained, we perform a set of recognition experiments. All scripts and result logs are included so that users of the toolbox can easily verify that they have successfully trained a set of acoustic models.

3.2.1. Recognition with syllable loop network

We first performed syllable decoding with a syllable loop word net. This recognition task puts the highest demand on the quality of the acoustic models. All 1679 syllables are listed in the network and any syllable can be followed by any other syllable, or they may be separated by short pause or silence.

The standard HTK decoder *HVite* was used for the experiment. A pruning beam width threshold of 120 was used to speed up recognition experiments.

3.2.2. Recognition with syllable bigram language model

The frequency of different tonal syllables can vary widely. For example, there are many syllables with the "neutral" tone which rarely occur in daily use. We have included a syllable bigram language model that had been estimated from the training set syllable transcription with the tool *HLstats*. Since the standard HTK decoder *HVite* does not support the use of trigram language models, we are using the bigram instead of the trigram language model for this baseline experiment.

3.3. Baseline reference results

The results from the baseline experiments are presented in Table 4. Since recognizing Chinese tones is a very difficult task, we have also calculated results that do not count tone misrecognitions as errors (shown in Table 4 as the Base Syllable result).

While better results have been achieved previously on this test set, those results were derived with larger training sets and additional input features. The baseline reference results presented are state of the art results using the standard approaches in large vocabulary speech recognition. With the provided corpus, there are some additional ways of improving recognition performance. For example, recognition accuracy on the tonal syllable recognition task can be improved by incorporating pitch into the feature vector [8]. Since many researchers in Mandarin speech recognition start their research program with a focus on the discrimination of tones, we hope that this corpus will be useful for testing various tone discrimination algorithms and for comparing results from different organizations.

For organizations which have much more data than what are provided in this toolbox, the testing set provided in the toolbox can be used as a benchmark for measuring acoustic model performance. Unlike the case for English speech recognition, currently there are no widely available large vocabulary trigram language models for Mandarin dictation tasks. Thus, it would be difficult to make a fair comparison of character accuracy on a particular set of acoustic waveforms. By focusing on the syllable recognition task, an objective assessment of acoustic model performance can be accomplished.

	Base Syllable %Corr	Tonal Syllable %Corr
Without LM	74.79%	51.21%
With Syllable Bigram LM	77.34%	67.55%

Table 4: Baseline syllable recognition results on the test set of 500 sentences from 25 male speakers.

4. SAPI 5.0 SDK

Another component of the toolbox is a copy of the SAPI 5.0 SDK, which was released for download on the web at http://www.microsoft.com/speech in October 2000. While

the SDK itself is available for download off the web, many students would find it difficult to download an over 100 MB file from many Chinese universities, especially when additional fees are charged for connecting to non-Chinese websites.

The SDK contains both a reference Mandarin speech recognition engine and a text to speech engine. The speech recognition engine is capable of large vocabulary continuous speech recognition and context free grammar (CFG) based recognition. The toolkit has been used successfully by many visiting students at Microsoft Research China and in a course taught to computer science students at Peking University. The students have created diverse prototypes such as a telephone based email reader, multi-modal smart phone user interfaces, and screen navigators for visually handicapped users. Another example of an SDK application is a multimodal conversational agent [9]. Figure 1 displays one of the prototypes that have been created using the SDK engine.

More importantly, the SAPI 5.0 SDK can be redistributed. Thus, if a research group creates specific applications that are useful (e.g. a speech based interface for blind users), the research group can provide their application along with the engine so that end users can benefit from the application. There are currently professors in China who are working with the SAPI 5.0 SDK for precisely these types of applications [10][11].

Figure 1: A multimodal smart phone prototype created using the SAPI 5.0 SDK.

5. Summary

As computers become more powerful, there is a strong and growing interest in the study and research of speech technologies in Chinese universities. This toolbox will provide professors and students interested in studying this field an opportunity to jumpstart their research, regardless of whether their subject of interest is in speech recognition research, dialog modeling, or multimodal user interface design. The database provided has been designed so that it is large enough to allow researchers and students to get started in fundamental acoustic research with widely available machines. It is hoped that the provided training and testing sets will also provide baseline results and benchmarks that will allow comparison of results across research groups. We also hope to initiate a trend where some standard "recipes" are made available utilizing standard corpora and publicly available programs so that reproducing and comparing results from different organizations become easier over time.

6. Acknowledgements

The authors wish to acknowledge Xue-Dong Huang, Hsiao-Wuen Hon, Yun-Cheng Ju, and Kai-Fu Lee for initiating the collection of the speech corpus.

7. References

[1] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P., "The HTK Book", http://htk.eng.cam.ac.uk.

[2] Kawahara, T., et. al., "Free software toolkit for Japanese large vocabulary continuous speech recognition", *Proc. 6th International Conference on Spoken Language Processing*, Beijing, 2000.

[3] Sjolander, K., and Beskow, J., "Wavesurfer – an open source speech tool", *Proc. 6th International Conference on Spoken Language Processing*, Beijing, 2000.

[4] 1997 Mandarin Broadcast News Speech, Linguistic Data Consortium, http://morph.ldc.upenn.edu/Catalog/LDC98S73.html.

[5] The 863 Project, Chinese Academy of Sciences.

[6] Bin, M., and Huo, Q., "Benchmark results of triphone-based acoustic modeling on HKU96 and HKU99 Putonghua corpora", *Proc. ISCSLP 2000*, Beijing, 2000.

[7] Wang, H. C., Seide, F., Tseng C. Y., and Lee, L. S., "MAT-2000 – design, collection, and validation of a Mandarin 2000-speaker telephone speech database", *Proc. 6th International Conference on Spoken Language Processing*, Beijing, 2000.

[8] Chang, E., Zhou, J., Di, S., Huang, C., and Lee, K. F, "Large vocabulary Mandarin speech recognition with different approaches in modeling tones", *Proc. 6th International Conference on Spoken Language Processing*, Beijing, 2000.

[9] Zhang, B., Cai, Q., Mao, J., Chang, E., and Guo, B., "Spoken dialog management as planning and acting under uncertainty", submitted to Eurospeech 2001.

[10] Luk, R., Yeung, D., Lu, Q., Leung, E., Li, S. Y., and Leung, F., "Digital Library Access for Chinese Visually Impaired", *Proc. of the fifth ACM conference on ACM 2000 digital libraries*, 2000.

[11] Zhu Xiaoyan, personal communication.