International Symposium on Chinese Spoken
Language Processing (ISCSLP 2000)
Fragrant Hill Hotel, Beijing
October 13-15, 2000

DEVELOP TELEPHONY SPEECH RECOGNITION SYSTEMS FOR REAL-WORLD APPLICATION

ZHANG Xiangdong, YUAN Baosheng, JIA Ying, TUO Lingyun, YAN Yonghong

Intel China Research Center, Beijin

Edward.zhang@intel.com； http://www.intel.com/research/mrl/people/zhang_x.htm

Abstract

This paper introduces our initial effort in building Mandarin acoustic model for Chinese stock information retrieval system based on Intel's LVCSR framework [1][2]. To build a robust and accurate system, a number of experiments were conducted to find the optimal parameters in various levels such as front-end feature, phonetic transcription, etc. We conducted comparison experiment to find the optimal configuration on the bandwidths for the telephony acoustic model in general. To build an accurate task-specific modeling, we introduce a hybrid context-dependent modeling of which the task-dependent training data and the task-independent one are treated differently in the modeling. The experiment result on two task-specific applications shows the proposed modeling can produce significant WER reduction. The telephone corpora were collected at ICRC to improve the robustness against both noise and channel effects.

1. INTRODUCTION

The Internet is becoming more and more powerful and ubiquitous, but when people cannot access computer like driving and walking, the telephone is ideal channel for access the information, and the most natural and convenient way for a user to communicate with the computer is by the speech.

For Mandarin telephone speech recognition, quite a few studies have been reported in the past. IBM built a HMM-based telephone speech recognition system. The lexicon contained 44,000 words. It used a large telephone speech database, referred to as "Mandarin call home database", to train the HMM models. The database contains 7.4-hour spontaneous speech recorded from international telephone calls. The word and syllable recognition error rate was 70.5% and 58.7% respectively [3]. A comparative study [4] using the Dragon system with fast adaptation and speaker normalization for compensating the channel and speaker variations was also conducted. The word error rate was improved 50% evaluated based on the same "Mandarin call home database". But, this performance is still far worse than that achieved in close-talk microphone speech recognition.

Currently, the recognition rate achieved in large-vocabulary telephone speech recognition is still too low to be practically useful. [8] But, for the special cases of small- and medium-size vocabulary, the telephone speech recognition performance is already good enough for the developing real world applications such as the Auto-attendant system and stock information retrieval system, etc. In some small- vocabulary cases, very good recognition performance can be achieved even without considering the channel and noise compensations. Actually, several useful applications have been developed in the past few years. In GALAXY system [5], a multilingual speech recognition system was implemented for on-line accessing information, such as weather information retrieval. It showed that a dialogue system, embedding with a speech recognizer, which recognizes several thousand words, could be a very useful task-dependent information inquiry system.

In this paper, an HMM-based Mandarin telephone speech recognition building process and related issues are discussed. The telephone corpora were collected at ICRC to improve the robustness against both noise and channel effects.

As real-time system has difficulty in Ceptral Mean Subtraction (CMS) calculation, we disable this feature in system training. For robust issue, we also remove the energy from training and testing feature, just use 12 MFCC+ 12 Δ MFCC + 12 ΔΔ MFCC, totally 36 dimensions feature for system building. For optimal system configuration, 15.9 k Gaussians used. The system vocabulary size used is about 2000 based on about 1020 listed stocks, alternative pronunciations and various abbreviation (symbols) of the stocks.

2. DATABASE AND MODELING ISSUES

It is well known that telephone speech recognition is a difficult task. Channel distortion and background noise interference is two main factors that degrade the recognition performance. The word error rate (WER) of telephone speech recognition system is usually much higher than that of a close-talk microphone speech recognition system. The main difficulty of telephone speech recognition is related to many calls being very noisy, as they often come from public phones, some with hand-free mode, some from GSM mobile phone that has poor SNR in general.

This paper reviews recent work done in Intel China Research Center (ICRC) addressing the above mentioned problems for robust telephone speech recognition in real-world application. Telephone speech recognition start at ICRC in late 1999 to extend existing technology [1][2] of continuous speech dictation to telephone applications.

2.1 Data Collection

Robustness is a very important requirement for successful speech recognition application.

Collected field data can be used to construct more robust HMMs. However, field data are usually unbalanced. A significant disparity in the amount of occurrences for each vocabulary exists [7]. Thus, general-purpose data that is phonetic balanced and field data must be jointly used to train accurate and robust HMM. By error analysis, we found there are several reason for recognition error: noise, strong accent, voice loudness, utterance, speaker's break, cough and second voice source. To increase robustness of our telephone speech recognition system, we implement an energy-based speech end-point detector. It makes the recognizer less sensitive to variations in Signal-Noise-Ratio (SNR).

The utterances are spoken by speech donator into telephone headsets and recorded directly in 8-bit mu-law digital form through analog connection to the usual switched telephone network. [10]

We collected data of over 3000 speakers, and over 800 hours. Among this, 150 hours are from GSM channel; others are from PSTN channel. The text script domain distributed in Name Dialing, Stock Information retrieval, Command & Control, Number and Digit String, meeting scheduling, Weather Information (location), Flight ticket booking and Currency exchange. The goal of ICRC data collection was to provide a basic set of common spoken material suitable for training and evaluation of speech recognition system for telephone-based applications, particularly those that use names, places, times and numbers in a Mandarin context. All the speech material was manually verified.

To covering a lager number of applications with Intel LVCSR technology, we use some telephone transferred speech data with phonetic balanced script in training the acoustic model. It is an important resource for us to do the real world telephone speech recognition application research and development.

In addition to the task specific telephone data we collect, we also build a general-purpose telephone data corpus based on 863 high quality close-talk microphone speech corpus. The methodology to build telephone data corpus from desktop data corpus are as following.

First, resample the speech data from 22.05Khz or 16Khz to 8Khz, The resampling process [6] are first to increase sampling rate by an integer factor, called interpolation, then applies an lowpass anti-aliasing FIR filter which is designed with a Kaiser window to the interpolated sequence, finally decreasing the sampling rate for a sequence, called decimation. In its filtering process, resample assumes the samples at times before and after the given samples in x are equal to zero.

Second, with a special utility we wrote, transferring the cleaned 8khz wave data through analog telephone line to get the general-purpose (with broad phone coverage) telephone data.

2.2 Feature Extraction

The feature vectors used in our experiments are formed of 36 coefficients. After pre-emphasizing with a filter of $1-0.97z^{-1}$ and 25 ms Hamming Windowed, the speech is transformed to 12 Mel-Scale Cepstrum Coefficient (MFCC) with frame shift every 10ms for both training and test data. First and second derivatives of the 12 coefficient vectors are estimated on 5 successive frames.

Calculate presence/absence of speech, FFT, mel-frequency energy bands, a cosine transform (resulting in the standard high-performance spectral parameters for speech recognition, known as Mel Frequency Cepstral Coefficients, or MFCC's),

Cepstral Mean Subtraction (CMS): The mean of the Cepstral vectors of the whole utterance is subtracted to the Cepstral coefficients of the feature vector. As CMS requires waiting until the end of the sentence before the recognition phase can start. So we disable CMS in both real-time system and model training process.

2.3 HMM Building

The telephone acoustic model building share the process of Intel LVCSR model training. [2]

We use semi-syllable as the Mandarin phonetic unit. The monophone number in the HMM models is 213, it includes 1 silence model, 1 tee model, 25 initial models, 186 tonal final models. We view tee model("sp") and silence model("sil") as context-independent models, except "sil" they cannot act as context of tri-phone.

Silence removal is a simple but effective way to improve recognition performance. Long period of silence has different environmental attributions from speech. Thus it affects the performance of Cepstrum mean subtraction. The technique is to remove surplus silence from the feature file and keep only a short period of silence (say, 7~8 frames). The new feature file is then being recognized.

The bootstrapping step initializes a set of three state, single-mixture, and monophone models. A phoneme-level transcript marking the phone boundaries for each training utterance is required at this stage. Experience has shown that bootstrapping is not a critical step. Subsequent training steps can compensate for poorly bootstrapped models. Bootstrapping consists of three steps:

Viterbi Training: The state means and variances of each model are initialized by uniformly dividing each training example (feature sequence) in time. Then the following procedure is performed iteratively. The training set is processed using Viterbi algorithm to find the best state sequence. The features assigned to each state during the Viterbi search are used to recalculate the state's means and variances. This is repeated for at most 10 iterations.

Forward-Backward Training: Model parameters from the Viterbi training step are re-estimated from the training data using the Baum-Welch recursion.

Model Formatting: The model files from forward-backward training (*.hmm) are combined and rewritten to the global HMM files. For phones that do not occur in the bootstrapping data set, models are created by copying those of acoustically similar phonemes.

The embedded monophone training step performs Baum-Welch model re-estimation using entire utterances rather than individual phones. This relaxes phone boundaries to move in order to better modeling the data. The number of iterations of embedded training is fixed at three. At this stage phoneme-level transcripts are required. However, phone boundary information is not used. Since precise phone boundaries are difficult and costly to obtain, it is usually possible to greatly increase the amount of training data at this stage. The transcript file format is modified slightly for efficiency. Phoneme indices are used instead of phoneme names.

During embedded training it is possible to define “tied states”, i.e., states whose parameters are shared across models. However, this feature is apparently not being used at this time.

Triphone Model Initialization & Re-estimation: The ICRC recognition system is based on triphone models. Triphone models are HMMs that represent phoneme in right and left context. For example, the model called “d-a4+r” represents the phoneme “a4” when it is preceded by the phoneme “d” and followed by the phoneme “r”. Within-word triphones do not span word boundaries. Crossword triphones do span word boundaries and also contain the within-word triphones. The triphone models are initialized by copying the associated monophone models. The Baum-Welch embedded training procedure is then applied to re-estimate the triphone model parameters. Again, only three iterations are applied.

Triphone Merging: The number of triphones is usually large. Moreover, some triphones may occur very infrequently (or not at all) in the training data. To ensure that all triphones are well trained, rare triphones are merged. The merging step is based on linguistic knowledge as well as frequency of occurrence. A language-dependent set of yes/no questions is used for every phoneme to construct a binary decision tree that determines which triphones can to be merged. At the root of the tree is the generic triphone (e.g., “*-a4+*”). All triphones appear at the leaves of the tree. The deeper the triphone, the less likely (linguistically) it is. The trees are then pruned based on the frequency of occurrence of the (merged) triphones. The pruning step may have to be performed several times, adjusting the pruning threshold until the proper balance between diversity of models and sufficiency of training data is achieved. A rule of thumb is that models should have at least 100 training examples per mixture. Finally, the merged models are reassembled.

Cluster Splitting / Mixture Creation: Prior to this step, all model states consisted of a single Gaussian probability density function. The final training step is to split each state pdf into several Gaussian mixtures. First, all the existing densities are split in two by perturbing the means and retraining using performs Baum-Welch model re-estimation. Then this procedure is repeated until the desired number of mixtures is achieved. This step must be monitored carefully to ensure that there is sufficient training data for each mixture. For our field HMM models, 1.3K states decision tree was built, and each states use 12 Gaussian mixtures. Totally, 15.9k Gaussians were used to describe the HMM models.

3. RECOGNITION EXPERIMENT AND RESULTS

We experimented two bootstrap approaches for Hidden Markov Models of phonetic units. First is to do downsampled wide-band 863 speech file which has manual labeling information, and train the individual models for each monophone with viterbi alignment and forward-backward training sequentially; Second is to do forced alignment with the first-shot model to the domain specific data (stock info) to give the labeling information, then apply the time labeled data to bootstrap training-viterbi and forward-backward training. The WER is equal in statistical variance. So the conclusion is the difference in bootstrap approach and data does not affect the performance dramatically.

The theoretical bandwidth for telephone speech signal is 300~3400Hz, while most fundamental frequency of speech signal is below 300Hz. So it is of our great interest to find the optimal band pass frequency for the front-end feature extraction. The experiments were conducted to find the optimal frequency range for the telephony speech ASR. The result shows lower frequency (60~300Hz) has important information for recognition. When training with 180~3400 filtered data, testing with 60~3400 filtered data, the error rate is very high (over 40%), it means that different front-end filtering can greatly impact the system performance.

Table 1 Performance comparison on various bandwidth settings

FILTER BANDWIDTH	300~3400	180~3400	60~3400	0~4000
Word Error Rate	11.8%	9.1%	7.2%	7.5%

There are a few issues using CW tri-phone model in a practical application. First, it costs more time for decoding. Second, it needs more training data to build a robust model. Third, when the training data contains significant amount task-independent data of which the text scripts are not related to the recognition vocabulary, many irrelevant CW triphones are introduced to the task. To address these issues, we use a hybrid approach in choosing the tri-phone units under the within-word (WW) modeling.

In the training data transcription, we treat every stock name as word, but treat every character (syllable) in the other task-independent text script (such as person’s name, entity’s name, word and phrases in dialog scenario and "863" sentences). As a result, cross-syllable tri-phone modeling is used for task relevant training data and on the other hand, within-syllable context-dependent modeling is used for task-independent training data. In this way, the task-specific models are modeled more accurately, and the task-independent ones are modeled more robust.

We built two task-specific acoustic models for two telephony applications, one is stock info model and the other is for Auto Attendant (AA) model of which the person's names are mainly in the task vocabulary. The impact of the task-specific word unit selection can be shown in the cross test for two task specific acoustic models.

We use same front-end stratagem and LVCSR framework, different word set to build the within word training transcript. Then do a cross testing. Stock model have 24.4% error rate increase on cross testing. AA model has 22% error rate increase on cross testing. In other word, for the same training data we can use task specific training script setting to favor the task specific system performance.

Table 2 Performance comparison on various task-dependent triphone model selection

TESTING SET	AA TEST DATA	STOCK TEST DATA
AA model	4.5%	8.8%
Stock model	5.6%	7.2%

4. CONCLUSIONS

This paper introduces our initial effort in building Mandarin acoustic model for Chinese stock information retrieval system. The telephone corpora were collected at ICRC to improve the robustness against both noise and channel effects. To build robust and accurate system, a number of experiments were conducted to find the optimal parameters in various levels such as front-end feature, phonetic transcription, etc. First experiment shows that different bootstrap approaches do not impact the system performance. The context triphone model embedded training steps play critical role in the system performance. Second comparison experiment is taken to find the optimal configuration on the bandwidths for the telephony acoustic model in general. The result shows lower frequency (60~300Hz) has important information for recognition. To build an accurate task-specific modeling, we introduce a hybrid context-dependent modeling of which the task-dependent training data and the task-independent one are treated differently in the modeling. The experiment result on two task-specific applications shows the proposed modeling can produce significant WER reduction.

5. REFERENCES:

[1] Yonghong. Yan, Xintian. Wu, J. Schalkwyk, R. Cole. " Development of CSLU LVCSR: The 1997 DARPA HUB4 Evaluation System”. In Proceedings DARPA '98 BNTUW, 1998.

[2] Intel Speech Development Toolkit (ISDT) Reference Manual, Intel Corp. 2000

[3] F.H. Liu, M. Picheny, P. Srinivasa, M. Monkowski, and J. Chen, "Speech Recognition on Mandarin Call Home: A Large Vocabulary, Conversational, and Telephone Speech Corpus", ICASSP'96, Vol. 1, pp. 157-160

[4] Barnett, J.; Corrada, A.; Gao, G.; Gillick, L.; Ito, Y.; Lowe, S.; Manganaro, L.; Peskin, B. " Multilingual speech recognition at Dragon Systems ", ICSLP 96. Vol4. Page(s): 2191 -2194

[5] Zue, V.; Seneff, S.; Polifroni, J.; Meng, H.; Glass, J. “ Multilingual human-computer interactions: from information access to language learning " ICSLP 96. Vol4, Page(s): 2207 -2210

[6] Crochiere, R., and L. R. Rabiner. 1983. Multirate Digital Signal Processing. Englewood Cliffs, NJ: Prentice-Hall, Inc.

[7] Mitchel Weintraub, Leonardo Neumeyer, Constructing Telephone Acoustic Models From A High-Quality Speech Corpus, ICASSP94,

[8] Yih-Ru Wang; Sin-Horng Chen “Mandarin telephone speech recognition for automatic telephone number directory service” ICASSP 98, vol.2, Page(s): 841 -844

[9] Mokbel, C.; Mauuary, L.; Jouvet, D.; Monne, J.; Sorin, C.; Simonin, J.; Bartkova, K. "Towards improving ASR robustness for PSN & GSM telephone applications ", Proceedings.of Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, 1996 , Page(s): 73 -76

[10] Bernstein, J.; Taussig, K.; Godfrey, J. " Macrophone: an American English telephone speech corpus for the Polyphone project ", ICASSP-94. Vol. 1 , Page(s): I/81 - I/84