10

Discrete-Mixture HMMs-based Approach for Noisy Speech Recognition

Tetsuo Kosaka, Masaharu Katoh and Masaki Kohda
Yamagata University
Japan

1. Introduction

It is well known that the application of hidden Markov models (HMMs) has led to a dramatic increase of the performance of automatic speech recognition in the 1980s and from that time onwards. In particular, large vocabulary continuous speech recognition (LVCSR) could be realized by using a recognition unit such as phones. A variety of speech characteristics can be modelled by using HMMs effectively. The HMM represents the transition of statistical characteristics by using the state sequence of a Markov chain. Each state of the chain is composed by either a discrete output probability or a continuous output probability distribution. In 1980s, discrete HMM was mainly used as an acoustic model of speech recognition. The SPHINX speech recognition system was developed by K.-F. Lee in the late 1980s (Lee & Hon, 1988). The system was a speaker-independent, continuous speech recognition system based on discrete HMMs. It was evaluated on the 997-word resource management task and obtained a word accuracy of 93% with a bigram language model. After that, comparative investigation between discrete HMM and continuous HMM had been made and then it was concluded that the performance of continuous-mixture HMM overcame that of discrete HMM. Then almost all of recent speech recognition systems use continuous-mixture HMMs (CHMMs) as acoustic models.

The parameters of CHMMs can be estimated efficiently under assumption of normal distribution. Meanwhile, the discrete Hidden Markov Models (DHMMs) based on vector quantization (VQ) have a problem that they are effected by quantization distortion. However, CHMMs may unfit to recognize noisy speech because of false assumption of normal distribution. The DHMMs can represent more complicated shapes and they are expected to be useful for noisy speech.

This chapter introduces new methods of noise robust speech recognition using discrete-mixture HMMs (DMHMMs) based on maximum *a posteriori* (MAP) estimation. The aim of this work is to develop robust speech recognition for adverse conditions which contain both stationary and non-stationary noise. Especially, we focus on the issue of impulsive noise which is a major problem in practical speech recognition system.

DMHMM is one type of DHMM frameworks. The method of DMHMM was originally proposed to reduce computation costs in decoding process (Takahashi et al., 1997).

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

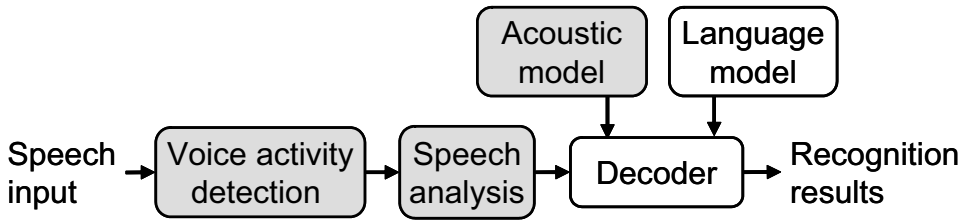


Figure 1. Block diagram of speech recognition system

DMHMM has more advantages of recognition performance than original DHMM. Nevertheless, the performance of DMHMM was lower than that of CHMM. In our work, we propose some new methods to improve the performance of DMHMM and demonstrate that the performance of DMHMM overcome that of CHMM in noisy conditions. We describe the characteristics of DMHMM and indicate the effectiveness and impact of the DMHMM for noisy speech recognition.

2. Noise Robust Speech Recognition

2.1 Basic approach of speech recognition under noisy conditions

Many efforts have been made for the issue of noise robust speech recognition over the years. For example, Parallel Model Combination (PMC) (Gales & Young, 1993), and Spectral Subtraction (SS) (Boll, 1979) are well known as effective methods of noisy speech recognition. And many other methods of noise robust speech recognition have been proposed. They are organized by some category. Fig. 1 is the block diagram of speech recognition system. Here, categorized methods of noise robust recognition are explained by using this figure. The noise robust methods can be roughly categorized into three groups: voice activity detection (VAD), speech analysis and acoustic modeling.

First, input speech is processed in voice activity detector. In this module, the presence or absence of speech is determined in noisy environment. If the speech detection fails, it is difficult to recognize input speech accurately. In recent years, the investigation on noise robust VAD has become active.

After a speech segment is detected, the speech signal is analyzed to extract the useful information for speech recognition. A cepstral analysis is a popular method for feature extraction in speech recognition. In particular, mel-scale frequency cepstrum coefficients (MFCC) are widely used as speech parameter in recent speech recognition system. Some noise reduction algorithms have been proposed to remove noise in speech waveform. The spectral subtraction method we mentioned above is one of those algorithms. One of the working groups in the European Telecommunication Standards Institute has approved the front-end (feature extraction module) for distributed speech recognition (ETSI, 2002). In this front-end, a noise reduction algorithm based on Wiener filter theory is employed. It is well known that this front-end is effective for various kinds of noise conditions and it is used as the baseline of the evaluation of proposed noise robust technique. Some new feature

extraction methods have been also proposed. Relative-Spectral Perceptual linear prediction (RASTA-PLP) method is one of the extraction methods and was reported that it was effective for noisy speech recognition (Hermansky et al., 1992).
Analysed speech is recognized in decoding process. In the decoder, a search algorithm is carried out with acoustic and language models. Those acoustic models are basically trained by clean data in which training utterances are recorded in quiet condition. However, models trained in clean condition cause a mismatch between models and input features in noisy condition. In order to eliminate it, many methods have been proposed. Multi-condition training is direct way to eliminate the mismatch condition (Pearce & Hirsch, 2000). In this training method, several types of noises are artificially added to the 'clean speech' data at several SNRs. The obtained noisy data were used for creating acoustic models. It was reported that this training method was effective, even if the noise conditions between training data and test data were different. Parallel Model Combination we mentioned above is also the one of the effective methods in this category. In this method, noisy speech model can be derived by combining the noise model and the clean speech model. Due to the assumption that the speech signal is affected by the additive noise in the linear spectral domain, cepstral parameters of the models are transformed to liner spectral domain for the combination. After the model combination, combined parameters are re-transformed to cepstral domain again.

2.2 Problem to be solved in this work

We categorized the previous works in the field of noise robust speech recognition and introduced some proposed methods in section 2.1. All the three approaches are important and many methods have been proposed. Most of these researches are, however, focused on stationary noise in which spectrum of noise signal is stationary in time domain. In contrast, speech recognition in non-stationary noise environments remains as a major problem. In practical speech recognition systems, the speech signals can be corrupted by transient noises created by tongue clicking, phone rings, door slams, or other environmental sources. These noises have a large variety of spectral features, onset time and amplitude, and a modeling of those features is difficult. Here, we call them 'impulsive noise'. The aim of this work is to develop robust speech recognition technology for adverse conditions which contain both stationary and non-stationary noise. In particular, we focus on the issue of impulsive noise.

2.3 Two types of approaches for noisy speech recognition

In order to solve the problem as shown in Section 2.2, we employ DMHMM as acoustic model. While three approaches are introduced in Fig. 1, the proposed method is categorized into the model-based approach. For model-based approach of robust speech recognition, we propose two different strategies.
In the first strategy, adverse conditions are represented by acoustic model. In this case, a large amount of training data and accurate acoustic models are required to present a variety of acoustic environments. This strategy is suitable for recognition in stationary or slow-varying noise conditions, because modeling of stationary noise is easier than that of non-stationary noise. In recent years, large speech corpora which contain much amount of training data are available. For ASR in stationary or slow-varying noise conditions, the effectiveness of multi-condition training has been reported (Pearce & Hirsch, 2000). For the

multi-condition training, a scheme of accurate modeling is needed because a large amount of noisy data created artificially is available.
In contrast, such training method is inadequate to recognize speech under impulsive noise. As mentioned above, impulsive noise has a large variety of features. However hard you may try to collect speech data in impulsive noise conditions, accurate modeling is very difficult. Then the second strategy is based on the idea that the corrupted frames are either neglected or treated carefully to reduce the adverse effect.
In order to achieve robust speech recognition in both stationary and impulsive noise conditions, we employ both strategies in this work. The concrete methods to realize both strategies are described in the next section. They are based on DMHMM framework which is one type of discrete HMM (DHMM). The method of DMHMM was originally proposed to reduce computation costs in decoding process. Two types of DMHMM systems have been proposed in recent years. One is subvector based quantization (Tsakalidis et al., 1999) and the other is scalar based quantization (Takahashi et al., 1997). In the former method, feature vectors are partitioned into subvectors, and then the subvectors are quantized by using separate codebooks. In the latter, each dimension of feature vectors is scalar-quantized. The quantization size can be reduced largely by partitioning feature vectors. For example, quantization size was reported as 2 to 5 bits in the former, and in the latter method, it was 4 to 6 bits. Because the quantization size is small, the DMHMM system has superior trainability in acoustic modeling.

2.4 MAP estimation and model compensation in DMHMM approach

As we mentioned in the previous sub-section, two kinds of strategies are employed to achieve robust speech recognition in both stationary and non-stationary noise conditions. In order to realize the first strategy, we propose a new modeling scheme of DMHMMs based on maximum *a posteriori* (MAP) estimate. For the second strategy, a method of compensating the observation probabilities of DMHMMs is proposed.
First, a new method of the MAP estimated DMHMMs for the first strategy is described below. In recent speech recognition systems, continuous-mixture HMMs (CHMMs) are generally used as acoustic models. It is well known that the CHMM system has an advantage in recognition performance over discrete HMM system. The parameters of CHMMs can be estimated efficiently under assumption of Gaussian distribution. However, CHMMs may unfit to recognize noisy speech because of false assumption of Gaussian distribution.
Fig. 2 shows an example of the discrete probability in DMHMM estimated by the method which is described below. The xy-plane represents the cepstrum (c1- c2) space and the z-axis represents the probability. The estimation was performed on noisy speech. It is obviously found that the shape of the distribution is not similar to that of the Gaussian distribution. As just described, discrete HMM can represent more complicated shapes and they are expected to be useful for noisy speech.
Considering the use of DHMMs, the insufficient performance is the major problem. The main reason why DHMMs show worse performance than CHMMs is because the accuracy of quantization in DHMM is insufficient. There is a trade off between quantization size and trainability. It is well known that reduction of quantization size of DHMMs leads to increase of quantization distortions, conversely, increase of quantization size leads to a lack of training data and poor estimation of parameters. As described above, the DMHMMs require

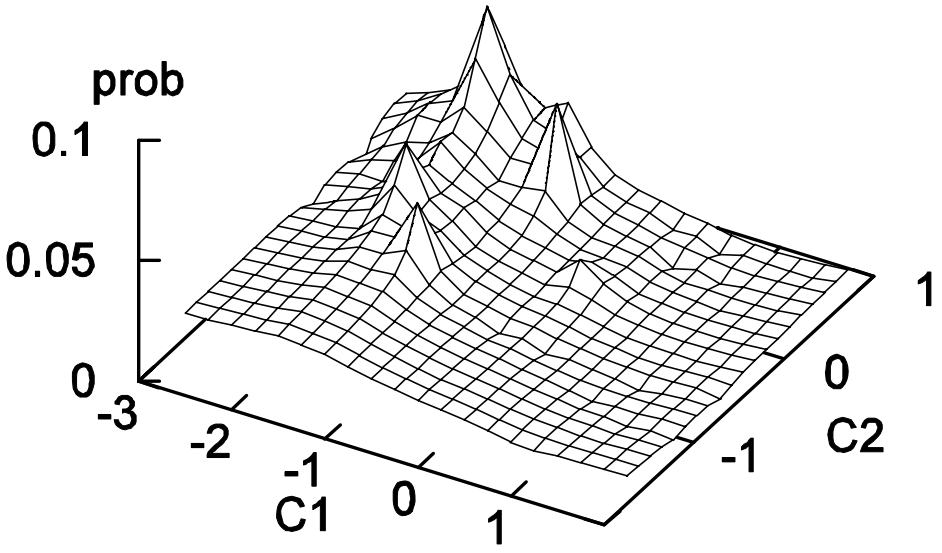


Figure 2. An example of discrete-mixture output distribution taken from the triphone 'a-a+a'.

a smaller amount of training data than ordinary discrete HMMs. Nevertheless, they still require larger amount of training data than CHMMs. In order to reduce the amount of training data and improve trainability further, we propose a new method for MAP estimation of DMHMM parameters. The MAP estimation is successfully used for adaptation of CHMM parameters (Lee & Gauvain, 1993). It uses information of an initial model as *a priori* knowledge to complement the training data.

In order to achieve the second strategy, we propose a method of compensating the observation probabilities of DMHMMs. Observation probabilities of impulsive noise tend to be much smaller than those of normal speech. The motivation in this approach is that flooring the observation probability reduces the adverse effect caused by impulsive noise. The method is based on missing feature theory (MFT) (Cooke et al., 1997) to reduce the effect of impulsive noise, so that the decoding process may become insensitive to distortions. In the MFT framework, input frames are partitioned into two disjoint parts, one having reliable frames and the other having unreliable frames which are corrupted by noise. Two different approaches are explored in the MFT framework: marginalization and imputation. In the marginalization approach, unreliable data are either ignored or treated carefully. The motivation of this approach is that unreliable components carry no information or even wrong information. In the imputation approach, values for the unreliable regions are estimated by knowledge of the reliable regions. The proposed compensation method is based on the first approach. Applying the MFT framework to speech recognition, it is difficult to determine reliable and unreliable regions. The proposed method does not require any determination of two regions in advance.

3. MAP Estimation of DMHMM Parameters

3.1 Discrete-Mixture HMMs

Before explaining MAP estimation of DMHMM parameters, the DMHMM proposed by Tskalidis (Tsakalidis et al., 1999) is briefly introduced here. As mentioned in Section 2, there are two types of DMHMM systems. In this paper, subvector based DMHMMs are employed. The feature vector is partitioned into S subvectors, $\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \cdots, \boldsymbol{o}_{st}, \cdots, \boldsymbol{o}_{St}]$. VQ codebooks are provided for each subvector, and then the feature vector $\boldsymbol{o}_t$ is quantized as follows:

$$q(\boldsymbol{o}_t) = [q_1(\boldsymbol{o}_{1t}), \cdots, q_s(\boldsymbol{o}_{st}), \cdots, q_S(\boldsymbol{o}_{St})]. \tag{1}$$

where $q_s(\boldsymbol{o}_{st})$ is the discrete symbol for the s-th subvector. The output distribution of DMHMM $b_i(\boldsymbol{o}_t)$ is given by:

$$b_i(\boldsymbol{o}_t) = \sum_m w_{im} \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) \tag{2}$$

$$\sum_m w_{im} = 1.0 \tag{3}$$

where w_{im} is the mixture coefficient for the m-th mixture in state i, and $\hat{p}_{sim}$ is the probability of the discrete symbol for the s-th subvector. In this equation, it is assumed that the different subvectors are conditionally independent in the given state and mixture index.

3.2 MAP Estimation

For creating HMMs from training data, maximum likelihood (ML) estimation is generally used as a parameter estimation method. MAP estimation is successfully used for adaptation of CHMM parameters (Lee & Gauvain, 1993). MAP estimation uses information from an initial model as *a priori* knowledge to complement the training data. This *a priori* knowledge is statistically combined with *a posteriori* knowledge derived from the training data. When the amount of training data is small, the estimates are tightly constrained by the *a priori* knowledge, and the estimation error is reduced. On the other hand, the availability of a large amount of training data decreases the constraints of the *a priori* knowledge, thus preventing loss of the *a posteriori* knowledge. Accordingly, MAP estimation tends to achieve better performance than ML estimation, if the amount of training data is small. The amount of training data tends to lack of the parameter estimation of DMHMMs, because the number of parameters in DMHMMs is larger than that in CHMMs.

In order to improve trainability further, we propose an estimation method of DMHMM parameters based on MAP. The ML estimate of discrete probability $p_{sim}(k)$ is calculated in the following form:

$$p_{sim}(k) = \frac{\sum_{t=1}^{T} \gamma_{imt} \delta(q_s(\boldsymbol{o}_{st}), k)}{\sum_{t=1}^{T} \gamma_{imt}} \tag{4}$$

$$\delta(q_s(\boldsymbol{o}_{st}), k) = \begin{cases} 1 & if \quad q_s(\boldsymbol{o}_{st}) = k \\ 0 & otherwise \end{cases} \tag{5}$$

where k is the index of subvector codebook and γ_{imt} is the probability of being in state i at time t with the m-th mixture component. Assume that the prior distribution is represented by Dirichlet distribution. The MAP estimate of DMHMM $\hat{p}_{sim}(k)$ is given by:

$$\hat{p}_{sim}(k) = \frac{(v_{simk} - 1) + n_{im} \cdot p_{sim}(k)}{\sum_{k=1}^{K}(v_{simk} - 1) + n_{im}} \tag{6}$$

where v_{simk} is the parameter of the prior distribution. And n_{im} is given as follows:

$$n_{im} = \sum_{t=1}^{T} \gamma_{imt} \tag{7}$$

In order to simplify the calculation, some constraints were added on the prior parameters. Assume that v_{simk} is given by:

$$v_{simk} = \tau \cdot p_{sim}^{0}(k) + 1 \tag{8}$$

$$\sum_{k=1}^{K} p_{sim}^{0}(k) = 1 \tag{9}$$

where $p_{sim}^{0}(k)$ is the constrained prior parameter. Applying Eq. (8) to Eq. (6), the MAP estimate $\hat{p}_{sim}(k)$ can be calculated by:

$$\hat{p}_{sim}(k) = \frac{\tau \cdot p_{sim}^{0}(k) + n_{im} \cdot p_{sim}(k)}{\tau + n_{im}} \tag{10}$$

where τ indicates the relative balance between the corresponding prior parameter and the observed data. In our experiments, τ was set to 10.0. Although both mixture coefficient and transition probability can be estimated by MAP, only output probability is estimated by MAP in this work.

3.3 Prior Distribution

The specification of the parameters of prior distributions is one of the key issues of MAP estimation. In this work, it is assumed that the prior distributions can be represented by models which are converted from CHMMs to DMHMMs. The conversion method is described below.

First, input vector $\boldsymbol{o}_t$ is divided into S subvectors, $\boldsymbol{o}_t = [\boldsymbol{o}_{1t}, \cdots, \boldsymbol{o}_{st}, \cdots, \boldsymbol{o}_{St}]$. The probability density of CHMMs in subvector s, state i and the m-th mixture component is given by:

$$b_{sim}^{'}(\boldsymbol{o}_{st}) = \frac{1}{(2\pi)^{\frac{d}{2}} \left|\boldsymbol{\Sigma}_{sim}\right|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{o}_{st} - \boldsymbol{\mu}_{sim})' \boldsymbol{\Sigma}_{sim}^{-1} (\boldsymbol{o}_{st} - \boldsymbol{\mu}_{sim})\right] \tag{11}$$

where $\boldsymbol{\mu}_{sim}$ is mean vector, $\boldsymbol{\Sigma}_{sim}$ is covariance matrix and d is the number of dimension for subvector s. In the case of $d=2$, $\boldsymbol{o}_{st}$, $\boldsymbol{\mu}_{sim}$ and $\boldsymbol{\Sigma}_{sim}$ are given by:

$$\boldsymbol{o}_{st} = \begin{bmatrix} o_{s_1 t} \\ o_{s_2 t} \end{bmatrix} \tag{12}$$

$$\boldsymbol{\mu}_{sim} = \begin{bmatrix} \mu_{s_1 im} \\ \mu_{s_2 im} \end{bmatrix} \tag{13}$$

$$\boldsymbol{\Sigma}_{sim} = \begin{bmatrix} \sigma^2_{s_1 im} & 0 \\ 0 & \sigma^2_{s_2 im} \end{bmatrix} \tag{14}$$

where s_1 and s_2 represent 1st and 2nd dimensions respectively. An output probability $b_i'(\boldsymbol{o}_t)$ is calculated by:

$$b_i'(\boldsymbol{o}_t) = \sum_m w_{im}' \prod_s b_{sim}'(\boldsymbol{o}_{st}) \tag{15}$$

where w_{im}' is the weighting coefficient of mixture component m. In order to obtain discrete parameters, the probability density for each centroid is calculated and normalized. As a result, parameters of the prior distribution $p^0_{sim}(k)$ are solved by:

$$p^0_{sim}(k) = \frac{b_{sim}'(\boldsymbol{\nu}_s(k))}{\sum_k b_{sim}'(\boldsymbol{\nu}_s(k))} \tag{16}$$

where $\boldsymbol{\nu}_s(k)$ is the centroid for each subvector s. While $p^0_{sim}(k)$ has a constraint of a normal distribution, $\hat{p}_{sim}(k)$ in Eq. (10) does not have such constraint. Thus it is expected that $\hat{p}_{sim}(k)$ will be updated to represent more complicated shapes in training session.

Some experimental results on MAP estimation are shown in Table 1 where word error rates (WERs) are indicated. 'DMHMM-ML' means ML estimated DMHMM and 'DMHMM-MAP' means MAP estimated DMHMM. MAP and ML estimation methods were compared at $SNR = \infty$ dB and $SNR = 5$ dB. The maximum number of training samples is 15,732. It is apparent that the performance of MAP is superior to that of ML. Then it can be concluded that the trainability of DMHMMs is improved by using MAP estimation.

SNR	model\number of samples	1000	5000	10000	15732
∞ dB	DMHMM-ML	100.0	54.66	27.54	18.63
	DMHMM-MAP	22.26	13.56	11.18	9.42
5dB	DMHMM-ML	99.0	45.01	31.99	29.45
	DMHMM-MAP	58.26	36.31	30.82	27.92

Table 1. Word error rate (%) results of the comparison between ML and MAP

4. Compensation of Discrete Distributions

In this section, a method of compensating the observation probabilities of DMHMMs is described to achieve robust speech recognition in noisy conditions. In particular, this method is effective for impulsive noise. The proposed method is based on the idea that the corrupted frames are either neglected or treated carefully to reduce the adverse effect. In other words, the decoding process becomes insensitive to distortions in the method. It is more likely that significant degradation of output probability appears in the case of mismatch conditions caused by unknown noise. Since the effect of impulsive noise is not considered in model training process, it is treated as unknown noise. If one of the subvector probabilities, $\hat{p}_{sim}(q_s(\boldsymbol{o}_{st}))$, is close to 0 in Eq. (2), the value of output probability, $b_i(\boldsymbol{o}_t)$, is also close to 0. It causes adverse effects in decoding process, even if the length of noise segment is short. Since an acoustic outlier such as impulsive noise is just unknown signal for acoustic model, difference in log likelihoods between outliers doesn't make sense for speech recognition. However, small difference between features of outliers causes large difference between log likelihoods, and it leads to changing the order of hypothesis in some situations.
In the proposed method, flooring the observation probability by threshold is employed. Since no difference in log likelihoods between outliers is shown in this method, it can reduce a negative effect in decoding process. Suppose that unreliable part can be found by using the value of discrete probability. In the proposed method, threshold for discrete probability is set, and negative effect is reduced in decoding process. Especially the method is effective for short duration noise. It is expected that pruning the correct candidate caused by impulsive noise is avoided.
For CHMM system, some compensation methods have been proposed with the same motivation. For example, Veth et al. proposed acoustic backing-off (Veth et al., 2001) where unreliable information is either neglected or treated carefully. Also the similar method was proposed in (Yamamoto et al., 2002). In those methods, the Gaussian distribution was compensated by threshold value. In our method, threshold can be set directly by value of probability. In other words, each threshold is given in the same way based on a probabilistic criterion. In contrast, it requires a kind of complicated way in CHMM system, because observation probabilities are given by probability density functions.
In the method proposed by Yamamoto (Yamamoto et al., 2002), single threshold was given for all Gussian distributions. In this case, since each shape of distribution is different, magnitude of the effect of threshold is also different. In the acoustic backing-off method, compensation values are different in each distribution. However, the compensation values depend on training data in the method. Comparison experiments with the acoustic backing-off are shown in Section 5.8.
Three types of compensation processing are proposed as follows:

Compensation at subvector level

A compensation is done at subvector level. In Eq. (2), $\hat{p}_{sim}(q_s(\boldsymbol{o}_{st}))$ is compensated by the threshold for subvector, *dth*.

$$\hat{p}'_{sim}(q_s(\boldsymbol{o}_{st})) = \begin{cases} \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) & if \quad \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) \geq dth \\ dth & otherwise \end{cases} \tag{17}$$

where $\hat{p}'_{sim}(q_s(\boldsymbol{o}_{st}))$ is the compensated discrete probability of $\boldsymbol{o}_{st}$. This threshold is especially effective in the case that specific subvector is corrupted.

Compensation at mixture level

A compensation is done at mixture level. In Eq. (2), $\prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st}))$ is compensated by the threshold for mixture component, *mth*.

$$\prod_s \hat{p}'_{sim}(q_s(\boldsymbol{o}_{st})) = \begin{cases} \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) & if \quad \prod_s \hat{p}_{sim}(q_s(\boldsymbol{o}_{st})) \geq mth \\ mth & otherwise \end{cases} \tag{18}$$

where $\prod_s \hat{p}'_{sim}(q_s(\boldsymbol{o}_{st}))$ is the compensated discrete probability of $\boldsymbol{o}_{st}$ at mixture level. This threshold is useful in the case that corruption affects a wide range of subvectors.

Compensation at both levels

Effectiveness of the above compensation methods depends on noise types. Thus, a combination of both compensation methods is expected to be effective for various types of noises.

Three types of compensation methods were compared in noisy speech recognition. The result of comparison in average error rate reduction among three was mixture level < subvector level < combination at both levels, and the reduction rate was 30.1%, 48.2% and 48.5%, respectively. The combination method obtained the best performance. From the viewpoint of the calculation cost, however, the compensation at subvector is not bad because the performance is similar. In the subvector method, the best performance can be obtained by the threshold from 2.0×10^{-3} to 5.0×10^{-3}. In the case that threshold is set to 5.0×10^{-3}, 68.6% of probability values in discrete distributions are floored. It turns out that a large proportion of probability values are useless for speech recognition.

5. Overview of Speech Recognition System Using DMHMMs

5.1 System Configuration

An experimental system of speech recognition for the study of DMHMMs has been developed. In this section, we describe the overview of the system. The recognition system makes use of a statistical speech recognition approach that uses DMHMM as the acoustic model and statistical language models such as word bigrams. This type of recognition system is called a large vocabulary continuous speech recognition (LVCSR) system. It can recognize more than several thousands of different words. Fig. 3 shows a block diagram of the system. It employs a time-synchronous beam search. The recognition results indicated in the previous sessions were obtained by this system.

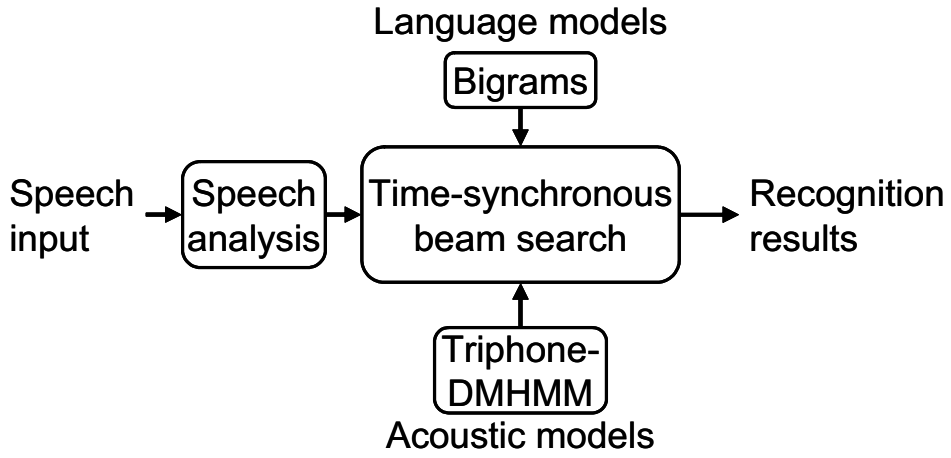


Figure 3. Speech recognition system

5.2 Speech Analysis

In the speech analysis module, a speech signal is digitized at a sampling frequency of 16kHz and at a quantization size of 16bits with the Hamming window. The length of the analysis frame is 32ms and the frame period is set to 8ms. The 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Additionally, the delta and delta-delta features are calculated from MFCC feature and log power. Then the total number of dimensions is 39. The 39-dimentional parameters are normalized by the cepstral mean normalization (CMN) method (Furui, 1974) which can reduce the adverse effect of channel distortion.

5.3 Decoder

For a large vocabulary continuous speech recognition (LVCSR), search space is very large and an expensive computation cost is required. A language model, which represents linguistic knowledge, is used to reduce search space. The detail of the language model is described in the next sub-section. In our system, one-pass frame-synchronous search algorithm with beam searching has been adopted. The searching algorithm calculates acoustic and language likelihood to obtain word sequence candidates. These word sequence candidates are pruned according to their likelihood values to reduce the calculation cost. Triphone models and word bigrams are used as acoustic and language models, respectively.

5.4 Language Model

A bigram is an occurrence probability of a pair of words that directly follow each other in text and is used as a linguistic constraint to reduce a calculation cost and improve recognition performance. The set of the probabilities are calculated with a large amount of text data. In this system, the bigrams have a 5000-vocabulary and are trained from 45 months' worth of issues of the Mainichi newspaper. Those data are in the database of 'JNAS: Japanese Newspaper Article Sentences'. It contains speech recordings and their

orthographic transcriptions. Text sets for reading were extracted from the articles of newspaper. The Mainichi Newspaper is one of the major nation-wide newspapers in Japan.

5.5 Acoustic Model

In recent years, context dependent models are widely used as acoustic models for speech recognition, since allophones or co-articulations can be modeled more accurately than context independent ones. Triphone model is one of the context dependent models and both the left and the right context are taken into consideration. It is well known that triphone is the effective model for continuous speech recognition. However, there is a problem when model parameters of triphone are estimated. The number of models exponentially increases depending on the number of contextual factors ant it causes the decrease of estimation accuracy. Then state sharing technique is widely used for context dependent models. In our system, shard-state triphone DMHMMs are uses as acoustic models. The topology of shard-state DMHMMs is represented by a hidden Markov network (HM-Net) which has been proposed by Takami (Takami & Sagayama, 1992). The HM-Net is a network efficiently representing context dependent left-to-right HMMs which have various state lengths and share their states each other. Each node of the network is corresponding to an HMM state and has following information:

- state number,
- acceptable context class,
- lists of preceding states and succeeding states,
- parameters of the output probability density distribution,
- state transition probabilities.

When the HM-Net is given, a model corresponding to a context can be derived by concatenating states which can accept the context from the starting node to the ending node.

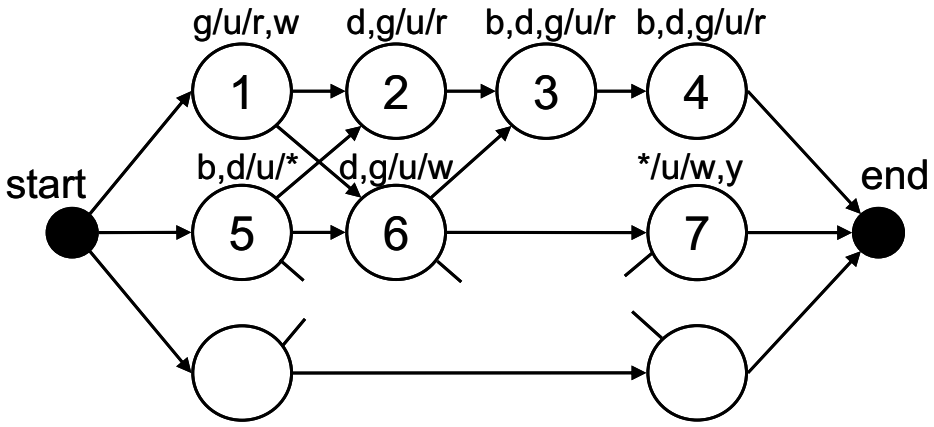


Figure 4. Example of the HM-Net

Fig. 4 shows an example of the HM-Net. In this figure, A/B/C stands for the acceptable context class, where A, B and C are the acceptable preceding, base and succeeding phone classes respectively. And the asterisk represents a class consisted of all phones. For example, the model for a context g/u/w is derived from a string of the states 1, 6 and 7. The extracted model is equivalent to general left-to-right HMMs. The structure of HM-Net we used in this work is determined by the state clustering-based method proposed by Hori (Hori et al., 1998). Although the output probability distribution of a HMM state is represented by Gaussian mixture density in original HM-Net, discrete mixture density is used in this system. The acoustic model we used is 2000-state HM-Net, and the number of mixture components was 4, 8 and 16, respectively. The 2000-state continuous mixture HM-Net is also prepared, and it is used for comparative experiments. As a comparative experiments in various number of mixture components, 16-mixture HM-Net shows the best result with either DMHMMs or CHMMs.

5.6 Codebook design of DMHMM

The codebook design in our experiments was determined in reference to the results in the paper written by Tsakalidis (Tsakalidis et al., 1999) and the split vector quantizer in the DSR front-end (ETSI, 2002). Tsakalidis has reported that DMHMMs with from 9 to 24 subvectors showed better performance. The feature vector is partitioned into subvectors that contain two consecutive coefficients. The consecutive coefficients that comprise subvector are expected to be correlated more closely. It was also reported that subvectors that contained consecutive coefficients performed well. Table 2 shows subvector allocation and codebook size. In the table, although delta and delta-delta parameters are omitted, those codebooks are designed in the same manner. The total number of codebooks is 21. The LBG algorithm was utilized for creating the codebook. Two types of codebooks were generated: 1) Clean codebook: A codebook derived from clean data. 2) Noisy codebook: A codebook derived from multi-condition data. Fig. 5 shows the examples of two codebooks. One represents c1-c2 plane, and the other is $\Delta c1-\Delta c2$ plane. Each point represents the codebook centroid and its number is 64 on c1-c2 or $\Delta c1-\Delta c2$ plane. Both clean codebook and noisy codebook are shown. As the experiment results, the performance of the DHMMs with noisy codebooks overcame that with clean codebooks for noisy speech recognition.

5.7 Training Data for Acoustic modeling

There are two sets of training data. They are on JNAS database. One is used for clean training, and the other is used for multi-condition training. The training data set consists of 15,732 Japanese sentences uttered by 102 male speakers. For clean training, no noise was added to the data. For multi-condition training, those utterances were divided into 20 subsets. No noise was added to 4 subsets. In the rest of the data, noise was artificially added.

parameter	logP, c0	c1,c2	c3,c4	c5,c6	c7,c8	c9,c10	c11,c12
codebook size	256	64	64	64	64	64	64

Table 2. Codebook design

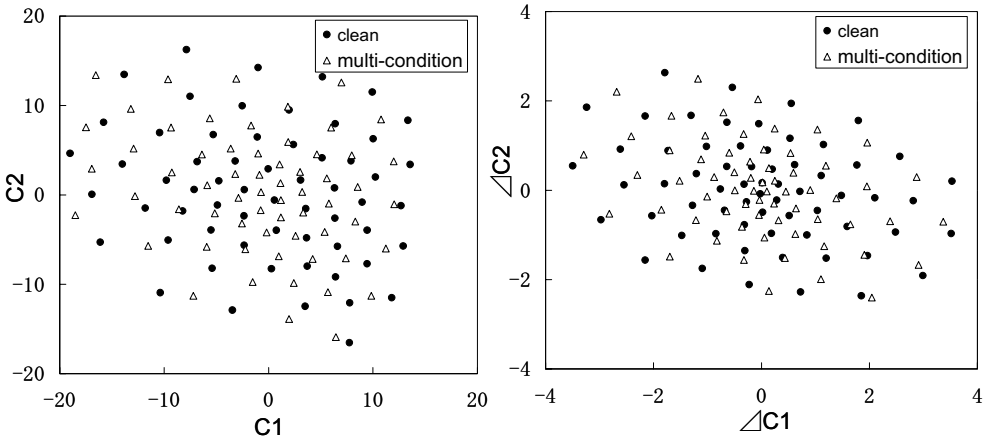


Figure 5. Examples of the codebooks (left figure: $c1-c2$ plane, right figure: $\Delta c1-\Delta c2$ plane)

Four types of noise (train, crowd, car and exhibition hall) were selected, and those were added to the utterances at SNRs of 20, 15, 10 and 5dB. The set of clean training data was used for parameter estimation of initial CHMMs and also used for clean codebook creation. The set of multi-condition training data was used for parameter re-estimation of both CHMMs and DMHMMs, and that was also used for creation of noisy codebook.

The procedures of acoustic model training are described as follows: First, initial CHMMs were trained by using clean speech data. Then the CHMMs were converted into DMHMMs using Eq. (16). The parameters of the DMHMMs derived here were used as the prior parameters of MAP estimation. After that, MAP estimation of DMHMMs was carried out by using multi-condition data.

5.8 Comparative experiments with other methods

Comparative experiments with other methods were conducted in adverse conditions that contain both stationary and impulsive noises. Noise signals from two test sets were mixed artificially to make new testset. Twelve types of noises (4 noise types (train, crowd, car and exhibition hall) times 3 SNR conditions) were prepared as stationary or slow-varying noise. These noises were mixed with 3 impulsive noises which were selected from RWCP database (Nakamura et al., 2000). Thus 36 types of noises were used for evaluation. Three types of impulsive noise were as follows:

- whistle3 blowing a whistle
- claps1 handclaps
- bank hitting a coin bank

The impulsive noises were added at intervals of 1-sec into speech data at SNR of 0 dB.

In the experiments, DMHMMs were compared with conventional CHMMs and CHMMs with acoustic backing-off method which is introduced in Section 4. As we described in the previous section, the compensation of discrete distributions is effective for noisy speech recognition. The acoustic backing-off is a similar method for CHMMs. In this method,

likelihood is calculated by using robust mixture distribution which consists of distribution $p(y)$ and 'outlier' distribution $p_o(y)$. Log likelihood $\log(p_{ab}(y))$ is given by

$$\log(p_{ab}(y)) = \log\{(1-\varepsilon)p(y) + \varepsilon \cdot p_o(y)\} \tag{19}$$

$$p_o(y) = (R_{\max} - R_{\min})^{-1} \tag{20}$$

where ε is the backing-off parameter, $R_{\max}$ and $R_{\min}$ are the maximum and minimum values for each component as observed in the training data.

Comparative experiments between acoustic backing-off and the proposed method were carried out. In the paper by Veth (Veth et al., 2001), $R_{\max}$ and $R_{\min}$ are given the maximum and minimum values of training data respectively. To avoid the dependence on training data, $R_{\max}$ and $R_{\min}$ are set to

$$R_{\max} = \mu + r \cdot \sigma^2 \tag{21}$$

and

$$R_{\min} = \mu - r \cdot \sigma^2 \tag{22}$$

where σ^2 is a variance of training data. Both r and ε were varied to find the best performance. As a result, r and ε were set to 3.0 and 5.0×10^{-5}, respectively.

The results of the comparison among DMHMMs, CHMMs and CHMMs with the acoustic backing-off are shown in Table 3. The threshold value of compensation for subvector and mixture component were set to 5.0×10^{-3} and 1.0×10^{-40}, respectively.

The proposed method shows the best performance among three methods. In the table, 'improvement' means the average error rate reduction from CHMM. It was 28.1% with the proposed method. In contrast, it was only 5.5% with the acoustic backing-off method. For DMHMMs, various thresholds for subvector were applied. The best performance was obtained by the threshold of 2.0×10^{-3}. More detailed results can be shown in the paper by Kosaka (Kosaka et al., 2005).

The results of CHMMs were too bad. It has been generally believed that the recognition error rates of DHMM were much higher than those of CHMM until now. Our experiments showed that the DMHMM framework performed better than conventional CHMM in noisy condition. In contrast, it was found that CHMM system showed similar or even better performance at high SNR in our experiments. In clean condition, the WER of DMHMMs was 6.7% and that of CHMMs with the acoustic backing-off was 6.6%. Recognition in clean conditions remains as an issue to be solved in the DMHMM system.

6. Conclusions

This chapter introduced a new method of robust speech recognition using discrete-mixture HMMs (DMHMMs) based on maximum *a posteriori* (MAP) estimation. The aim of this work was to develop robust speech recognition for adverse conditions which contain both stationary and non-stationary noise. In order to achieve the goal, we proposed two methods.

First, an estimation method of DMHMM parameters based on MAP was proposed. The second was a method of compensating the observation probabilities of DMHMMs to reduce adverse effect of outlier values. Experimental evaluations were done on Japanese speech recognition for read newspaper task. Compared with conventional CHMMs and CHMMs using the acoustic backing-off method, MAP estimated DMHMMs performed better in noisy conditions than those systems. The average error rate reduction from CHMMs was 28.1% with the proposed method. It has been generally believed that the recognition error rates of DHMM were much higher than those of CHMM until now. However, our experiments showed that the DMHMM framework performed better in noisy conditions than conventional CHMM framework.

method\noise		WER(%)			improvement(%)
		whistle3	claps1	bank	
CHMM		65.9	43.9	37.5	-
CHMM-AB		65.2	39.5	36.8	5.5
DMHMM	5.0×10^{-4}	51.5	34.9	31.2	22.7
	2.0×10^{-3}	46.8	32.9	30.5	28.1
	5.0×10^{-3}	46.5	35.2	31.2	24.7

Table 3. The WER results of the comparison among three methods in mixed noise conditions

Although, the proposed method is effective in noisy conditions, its performance is insufficient in clean conditions. What we are aiming for as a future work is to improve trainability and recognition performance in clean conditions further.

7. Future prospects

We are now conducting the evaluation of the method on more difficult task. In Japan, a large-scale spontaneous speech database 'Corpus of Spontaneous Japanese (CSJ)" has been used as the common evaluation database for spontaneous speech now (Furui et al., 2005). This corpus consists of roughly 7M words with a total speech length of 650 h. In the corpus, monologues such as academic presentations and extemporaneous presentations have been recorded. The recordings were carried out by a headset microphone with relatively little background noise. It is well known that the recognition of this task is too difficult because those presentations are real and the spontaneity is high. For example, 25.3% of word error rate was reported by Furui (Furui et al., 2005). In our experiments, 20.72% of word error rate has been obtained with 6000-state 16-mixture DMHMMs and the trigram model of 47,099 word-pronunciation entries (Yamamoto et al., 2006). It shows that DMHMM system has a high performance even if in low noise conditions. The DMHMM-based system has much more potential for speech recognition, because it needs no assumption of Gaussian

distribution. For example, model adaptation in which the shape of distribution of HMM is modified intricately cannot be carried out in CHMM framework, but could be done in DMHMM. We plan to develop DMHMM-related technologies further for improving speech recognition performance.

8. References

Boll, S. (1979), Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. on Acoustics, Speech and Signal Processing,* Vol. ASSP-27, No. 2, Apr. 1979, pp.113-120

Cooke, M. P.; Morris, A. & Green, P. D. (1997), Missing data techniques for robust speech recognition, *Proceedings of ICASSP97,* pp.863-866, Munich, Germany, Apr. 1997, IEEE

ETSI, (2002), ETSI ES 202 050 V1.1.1, *STQ; Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms,* European Telecommunications Standards Institute, France

Furui, S. (1974), Cepstral analysis technique for automatic speaker verification, *Journal of Acoustical Society of America,* Vol. 55, Jun. 1974, pp. 1204-1312

Furui, S.; Nakamura, M.; Ichiba, T. & Iwano, K. (2005), Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese, *Speech Communication*, Vol. 47, Sept. 2005, pp. 208-219, ISSN 0167-6393

Gales, M & Young, S. (1993). HMM recognition in noise using parallel model combination, *Proceedings of Eurospeech 93,* pp. 837-840, Berlin, Germany, Sept. 1993, ESCA

Hermansky, H.; Morgan, N.; Baya, A. & Kohn, P. (1992), RASTA-PLP speech analysis technique", *Proceedings of ICASSP92*, pp. 121-124, Sam Francisco, USA, Mar. 1992, IEEE

Hori, T.; Katoh, M.; Ito, A. & Kohda M. (1998), A study on a state clustering-based topology design method for HM-Nets, *IEICE Transactions (Japanese)*, Vol. J81-D-II, No. 10, Oct. 1998, pp. 2239-2248

Kosaka, T.; Katoh, M. & Kohda M. (2005), Robust speech recognition using discrete-mixture HMMs, *IEICE Transactions*, Vol. E88-D, No. 12, Dec. 2005, pp. 2811-2818

Lee, C.-H. & Gauvain, J.-L. (1993), Speaker adaptation based on MAP estimation of HMM parameters, *Proceedings of ICASSP93,* pp. 558-561, Minneapolis, USA, Apr. 1993, IEEE

Lee, K.-F. & Hon, H.-W. (1988), Large-vocabulary speaker-independent continuous speech recognition using HMM, *Proceedings of ICASSP88*, pp. 123-126, New York, USA, Apr. 1988, IEEE

Nakamura, S.; Hiyane, K.; Asano, F.; Nishimura, T. & Yamada, T. (2000), Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, *Proceedings of LREC2000*, pp. 965-968, Athens, Greece, May. 2000

Pearce, D & Hirsch, H.-G. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions, *Proceedings of ICSLP2000,* Vol.4, pp.29-32, Beijing, China, Oct. 2000

Takahashi, S.; Aikawa K. & Sagayama S. (1997). Discrete mixture HMM, *Proceedings of ICASSP97,* pp.971-974, Munich, Germany, Apr. 1997, IEEE

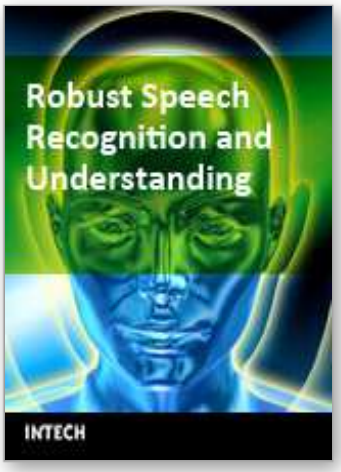
Takami, J. & Sagayama S. (1992). A successive state splitting algorithm for efficient allophone modeling, *Proceedings of ICASSP92*, pp. 573-576, San Francisco, USA, Mar. 1992, IEEE

Tsakalidis, S.; Digalakis, V. & Newmeyer, L. (1999). Efficient speech recognition using subvector quantization and discrete-mixture HMMs, *Proceedings of ICASSP99,* pp.569-572, Phoenix, USA, Mar. 1999, IEEE

Veth, J.; Cranen, B. & Boves, L. (2001), Acoustic backing-off as an implementation of missing feature theory, *Speech Communication*, Vol. 34, No. 3, Jun. 2001, pp. 247-265, ISSN0167-6393

Yamamoto, A.; Kumakura, T.; Katoh, M.; Kosaka, T. & Kohda, M. (2006), Lecture speech recognition by using codebook adaptation of discrete-mixture HMMs, *Proceedings of ASJ Autumn Meeting (Japanese)*, pp. 69-70, Kanazawa, Japan, Sept. 2006, ASJ

Yamamoto, H.; Shinoda, K. & Sagayama, S. (2002), Compensated Gaussian distribution for robust speech recognition against non-stationary noise, *Technical Report of IEICE (Japanese),* SP2002-45, pp.19-24, Sendai, Japan, Jun. 2002, IEICE



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821