Multi-Lingual Speech Recognition System for Speech-To-Speech Translation

Satoshi Nakamura, Konstantin Markov, Takatoshi Jitsuhiro,
Jin-Song Zhang, Hirofumi Yamamoto, Genichiro Kikui

Spoken Language Translation Research Labs.
Advanced Telecommunication Research Institute, Kyoto, Japan
{satoshi.nakamura,konstantin.markov,takatoshi.jitsuhiro}@atr.jp,
{jinsong.zhang,hirofumi.yamamoto,genichiro.kikui}@atr.jp

Abstract

This paper describes the speech recognition module of the speech-to-speech translation system being currently developed at ATR. It is a multi-lingual large vocabulary continuous speech recognition system supporting Japanese, English and Chinese languages. A corpus-based statistical approach was adopted for the system design. The database we collected consists of more than 600 000 sentences covering broad range of travel related conversations in each of the three languages. The recognition system is based on language-dependent acoustic and language models, and pronunciation dictionaries. The models are built using the latest training methods developed at ATR as the Minimum Description Length Successive State Splitting (MDL-SSS) and Multi-dimensional Composite N-gram techniques. The specifics of each language are taken into account in order to achieve high recognition performance. The speech recognition system is under constant improvement and enhancement, and although the models for the different languages are at different development stages, the recent evaluation experiments showed that the recognition performance is above 92% for every language.

1. Introduction

Speech-to-Speech (S2S) translation is a pipe dream for human-beings, which enables communication with people speaking in different languages. Since our world is becoming borderless day by day, the importance of S2S translation technology has been increasing. ATR had started the S2S translation research in order to overcome this language barrier problem in 1986. So far, we have been working in speech recognition, machine translation, speech synthesis and integration for a S2S translation system. We are currently in the third term beginning from 2000. The target of this term is to develop technologies to make the S2S system work in real environments. Speech recognition system should be robust enough to recognize speech in noisy environments with various speaking styles. The machine translation system needs to be domain portable and good to translate wide variety of topics. The speech synthesis is requested to realize more natural and expressional speech quality. In this project all the researchers including speech processing researchers and natural language researchers are working collaboratively and closely for realization of S2S translation system.

For the S2S system the speech recognition system should recognize speaker independent continuous spontaneous conversational speech. Back in 1986, the state of the art technology of speech recognition is only able to recognize speaker dependent connected words of small vocabulary. Due to a lot of efforts so far based on statistical modeling technologies like HMMs and N-grams and large amounts of speech and text corpus, now recognition of speaker independent continuous conversational speech is going to be available. The thing to consider especially developing speech recognition for S2S translation system is nothing but speech recognition system to be multi-lingual.

In speech recognition, it is well known that probabilities by acoustic modeling by HMMs and language modeling by N-gram are jointly used to search an optimal word sequence in decoding. Parameters of HMMs and N-grams are estimated using a large amount of speech data and text data. There is always trade-off between number of parameters for the model and number of training data. It always takes lots of time to find a best number of parameters suitable for the data amount. This paper uses new acoustic modeling procedure to create variable state assignment of HMM states based on successive state splitting and information criteria. The successive state splitting algorithm, SSS, is a procedure to split a existing state and assign variable number of states for one phoneme unit considering likelihood increase. Furthermore, MDL-SSS, which is SSS algorithm based on minimum description length, is proposed in order to stop growth of number of states according the training data amount. In multi-lingual situations the available amounts of training data are often different and this makes optimization process complicated. The MDL-SSS is proposed to solve this problem. The same problem also exists in language modeling. There are also an out-of-vocabulary problem and a new words problem. For this problems we introduce class-based N-gram and composite bigram to approximate trigram. It is also true that size of training text corpus amount is always different depending on language. An efficient estimation algorithm always help to get good estimate of language model parameters.

In this paper, we describe designs and algorithms of our multi-lingual speech recognition system for S2S translation system. Currently our target languages are Japanese, English and Chinese. This paper also introduces language dependent parts and their speech recognition evaluation results of each language as well.

2. Speech-to-Speech Translation Background

The goal of the automatic speech-to-speech translation is to generate a speech signal in one (target) language that conveys the linguistic information contained in a given speech signal of another (source) language.

A statistical approach to the speech-to-speech translation problem gives the following formal solution:

$$S_T^* = \arg\max_{S_T} P(S_T|S_S) \tag{1}$$

where S_S and S_T are the speech signals in the source and target languages. As direct evaluation of the conditional probability $P(S_T|S_S)$ is intractable, it can be factorized as:

$$\begin{aligned} P(S_T|S_S) &= \sum_{T_T,T_S} P(S_T,T_T,T_S|S_S) \\ &= \sum_{T_T,T_S} P(S_T|T_T,T_S,S_S)P(T_T|T_S,S_S)P(T_S|S_S) \\ &\approx \sum_{T_T,T_S} P(S_T|T_T)P(T_T|T_S)P(T_S|S_S) \end{aligned} \tag{2}$$

where T_S and T_T are the text transcriptions of the source and target speech signals. Then, the maximization of $P(S_T|S_S)$ can be further simplified to:

$$\begin{aligned} \max_{S_T} P(S_T|S_S) \quad = \quad & \max_{S_T} P(S_T|T_T^*) \max_{T_T} P(T_T|T_S^*) \\ & \max_{T_S} P(T_S|S_S) \end{aligned} \tag{3}$$

where T_T^* and T_S^* are arguments maximizing the second and third terms. This equation suggests that the S2S translation problem can be decomposed into three independent parts: $P(T_S|S_S)$ which represents speech recognition, $P(T_T|T_S)$ that is text-to-text translation model, and $P(S_T|T_T)$ which corresponds to speech synthesis.

For the speech recognition problem, suppose the input speech signal S_S is represented by a sequence of feature vectors $\boldsymbol{y}$:

$$\boldsymbol{y} = y_1 \cdots y_t \tag{4}$$

Then, the speech recognition goal is to find a word sequence $\boldsymbol{w} = w_1 \cdots w_n$ that maximizes $P(\boldsymbol{w}|\boldsymbol{y})$:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \underset{\boldsymbol{w}}{argmax}\, P(\boldsymbol{w}|\boldsymbol{y}) \\ &= \underset{\boldsymbol{w}}{argmax}\, P(w_1 \cdots w_n|y_1 \cdots y_t) \end{aligned} \tag{5}$$

This can be rewritten by Bayes theorem as:

$$P(\boldsymbol{w}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{y})} \tag{6}$$

here, $P(\boldsymbol{y})$ is a prior probability of the input speech which is independent on the word hypothesis. Then, the solution will be to search for $\boldsymbol{w}$ that maximizes, $P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})$. Here, $P(\boldsymbol{w})$ is *a priori* probability for the word sequence independent on the input speech signal. The model for $P(\boldsymbol{w})$ is called language model while the $P(\boldsymbol{y}|\boldsymbol{w})$ is represented by the acoustic model.

3. SSS-based Acoustic Modelling

For acoustic modeling in speech recognition, contextdependent phone models can obtain much better performance than context-independent phone models. While context-dependent phone models have many parameters, the most important problem to solve has been how to efficiently capture contextual and temporal variations in training speech and properly model them with fewer parameters.

Phonetic decision tree clustering[1] was proposed as a method for generating tied-state structures of acoustic models for speech recognition. The Successive State Splitting (SSS) algorithm was originally proposed to create a network of HMM states of speaker dependent models by ATR[2]. Furthermore, it was subsequently expanded to the ML-SSS algorithm to create speaker independent models[3] by data-driven clustering with contextual information.

However, since these methods base on the Maximum Likelihood (ML) criterion, the likelihood value for training data increases as the number of parameters increases. To overcome this problem, information criteria have been introduced for splitting and stop criteria to create tied state HMMs. In [4], a method using Akaike's Information Criterion (AIC) was proposed to determine the topologies of context-independent models. To create context-dependent models using phonetic decision tree clustering, the Minimum Description Length (MDL) criterion[5] and the Bayesian Information Criterion (BIC)[6][7] have been introduced.

In the SSS algorithm, this problem is more serious than in the decision tree clustering because the SSS algorithm can make models with more degrees of freedom. Therefore, we have recently introduced the MDL criterion into the ML-SSS algorithm to use it as the splitting and stop criteria[8]. This algorithm is referred to as "the

MDL-SSS algorithm.” We will describe the ML-SSS algorithm and the MDL-SSS algorithm in the following sections.

3.1. ML-SSS Algorithm

The ML-SSS algorithm iteratively constructs the appropriate context-dependent model topologies by finding a state which should be split in each iteration and then it re-estimates the parameters of HMMs based on the ML criterion in the same way as in phonetic decision tree clustering. This algorithm supposes the two types of splitting shown in Fig. 1.

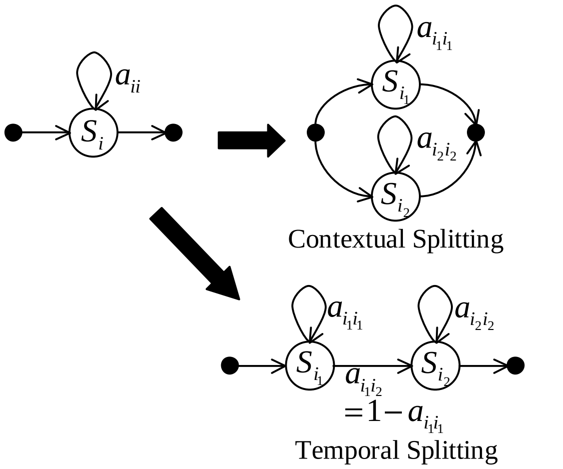


Figure 1: Contextual splitting and temporal splitting

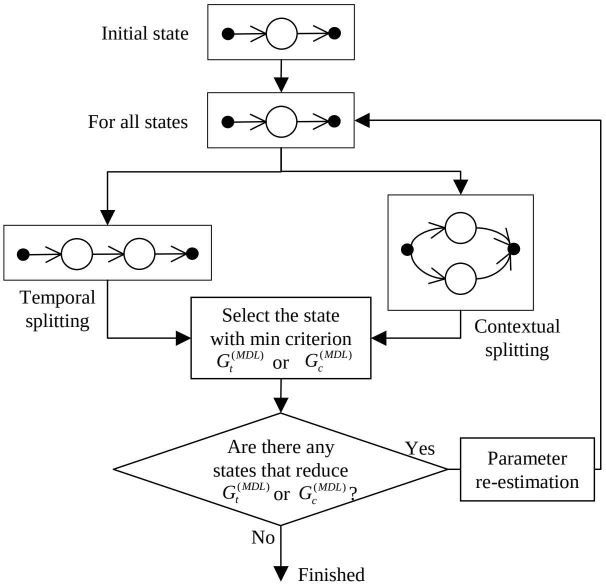
After contextual splitting, the total expected gain is calculated. For temporal splitting, the ML-SSS algorithm creates one more state and connects it to the original state. The parameters of the two distributions are estimated by the forward-backward algorithm, and the total expected gain of temporal splitting is also calculated for the temporal split states. Since it is computationally expensive to re-estimate all of the parameters of a network at every splitting, approximated likelihood values are used. Next, the gains of both contextual and temporal splitting are calculated for all states. Finally, these expected gains are compared with each other and the split with the best gain among all states is selected. N_s is the total number of states and N_p is the maximum temporal length of states for each triphone model. These parameters are stop criteria and must be given before starting the splitting. Nonetheless, it is difficult to find the optimal values of these parameters. Accordingly, a sequence of experiments needs to be done to find the optimal values by changing parameters heuristically.

3.2. SSS Algorithm Using MDL Criterion

Next, we introduce the MDL criterion to the ML-SSS algorithm. In this section, we define the MDL-SSS algorithm, which uses the MDL criterion instead of the ML criterion as the splitting criterion for the ML-SSS algorithm.

3.2.1. *Flow of MDL-SSS algorithm*

Figure 2 shows the flow of the MDL-SSS algorithm. The differences in the MDL values for both contextual and temporal splitting are calculated for each state, and the split with the smallest difference value is chosen. Splitting is finished when there is no state that can be split and reduce the criterion by splitting. The total number of states and the maximum number of states per triphone are not required as stop criteria.



Initial state
For all states
Temporal splitting
Contextual splitting
Select the state with min criterion $G_t^{(MDL)}$ or $G_c^{(MDL)}$
Are there any states that reduce $G_t^{(MDL)}$ or $G_c^{(MDL)}$?
Yes
No
Parameter re-estimation
Finished

Figure 2: Flow chart of MDL-SSS algorithm.

3.2.2. *Gain function by MDL-SSS*

We define the criteria for contextual splitting and temporal splitting, $G_c^{(MDL)}$ and $G_t^{(MDL)}$, respectively, as follows:

$$G_c^{(MDL)}(S_i) = -G_c^{(ML)}(S_i) + C_c K \log \Gamma(S), \quad (7)$$

$$\begin{aligned} G_t^{(MDL)}(S_i) &= -G_t^{(ML)}(S_i) \\ &\quad + C_t \frac{(2K+1)}{2} \{(M+1) \log \Gamma'(S) \\ &\quad - M \log \Gamma(S)\}, \end{aligned} \quad (8)$$

where the order of features is K, the total number of states is M. The first terms, $G_c^{(ML)}$ and $G_t^{(ML)}$, in the right-hand sides are the negative values of the expected gains in the ML-SSS algorithm. C_c and C_t are the scaling factors of the second terms. $\Gamma(S) = \sum_{i=1}^{N_s} \Gamma(S_i)$ represents the expected frequency of the number of samples for all states. $\Gamma'(S)$ is the value after temporal splitting.

Equation (8) compensates the total number of samples, $\Gamma(S)$, because segments that are shorter than the lengths of state sequences are discarded. $\Gamma(S)$ will be decreased to $\Gamma'(S)$ if a temporal split is selected. The MDL-SSS algorithm selects the state with the smallest $G_c^{(MDL)}$ or $G_t^{(MDL)}$, and stops splitting when $G_c^{(MDL)} > 0$ and $G_t^{(MDL)} > 0$ for all states.

4. Advanced Language Modelling

4.1. Multi-Dimensional Word Classes

In the conventional word class N-gram defined by:

$$P(w_i|w_{i-N+1}\dots w_{i-1}) = P(C(w_i)|C(w_{i-N+1})\dots C(w_{i-1}))P(w_i|C(w_i)) \quad (9)$$

only one-dimensional word classes are used. Both the left and right context Markovian dependencies are used together. Only the words having the same left and right context Markovian dependence belong to the same word class. This word class definition is not adequate to represent the Markovian dependence for words that have only the same left or right context Markovian dependence, such as ”a” and ”an”. The left context of ”a” and ”an” is almost equivalent, however the right context is significantly different. The difference between left and right context is more serious in languages with inflection, such as French and Japanese. For example, the Japanese inflection form has an influence only on the right context, the left context Markovian dependence can be shared between the same words with different inflection forms.

We have introduced the idea of Multi-dimensional word classes to represent left and right context Markovian dependence separately [10]. The multi-dimensional word classes can assign the same word class to ”a” and ”an” to represent the left context Markovian dependence (left context class), and assign them to different word classes to represent the right context Markovian dependence (right context class). Each multi-dimensional word class is automatically extracted from the corpus using statistical information, rather than grammatical information such as POS.

4.2. Class N-grams based on Multi-Dimensional Word Class

Applying multi-dimensional word classification to formula (9), the next formula is obtained.

$$P(w_i|w_{i-N+1}\dots w_{i-1}) = P(C^l(w_i)|C^{rN-1}(w_{i-N+1})\dots C^{r2}(w_{i-2}C^{r1}(w_{i-1})) P(w_i|C^l(w_i)) \quad (10)$$

where the suffix for class C is used to represent position dependent (left and right context) Markovian dependence. $C^l(w)$ represents the left context class to which the word w belongs, and $C^{ri}(w)$ represents the right context class to which the i-th word w belongs. Hereafter, we refer to these class N-grams based on multi-dimensional classes as Multi-Class N-grams.

4.3. Word Clustering for Multi-Class Bi-grams

For clustering, we adopt vectors to represent left and right context Markovian dependence, i.e., which words will appear in a left or right context with what probability. These Markovian dependence vectors are defined as follows:

$$v^l(x) = [P^b(w_1|x), P^b(w_2|x), ..., P^b(w_V|x)] \quad (11)$$

$$v^r(y) = [P^f(w_1|y), P^f(w_2|y), ..., P^f(w_V|y)] \quad (12)$$

where $v^l(x)$ represents the left context Markovian dependence vector of x. This vector is used for left context class clustering. $P^b(w_i|x)$ is the value of the probability of the backward bi-gram from x to w_i (i-th word in the lexicon). $v^r(y)$ represents the right context Markovian dependence vector of y. This vector is used for right context class clustering. $P^f(w_i|y)$ is the value of the probability of the forward bi-gram from y to w_i. V is the size of vocabulary.

For clustering, the distance between the word pair's vectors is used, since word pairs with similar vectors also have similar Markovian dependence. We use Euclidean distance as a distance measure. Word clustering is thus performed in the following manner, called the uni-gram weighted Ward method.

4.4. Use of Frequent Word Successions

Furthermore, Multi-class N-grams are extended to Multi-class composite N-grams. In this model, higher order word N-grams are partially introduced by regarding frequent variable length word sequences as new word succession entries. In this way, for frequent word sequences with length L, an L order word N-gram can be estimated reliably, even if the training corpus size is insufficient to estimate N-grams of other words. After introduction of higher order word N-grams, the increase of parameters only corresponds to a uni-gram of word succession. Therefore, Multi-class composite N-grams can maintain a compact model size in multi-class N-grams.

5. BTEC Description

The Basic Travel Expression Corpus (BTEC) was planned to cover utterances for every potential subject in travel conversations, together with their translations. Since it is almost infeasible to collect them through transcribing actual conversations or simulated dialogs, we decided to use sentences from the memories of bilingual travel experts. We started by investigating phrasebooks that contain Japanese/English sentence pairs that those experts consider useful for tourists traveling abroad. We collected these sentence pairs and rewrote them to make translations as context-independent as possible and to comply with our transcription style. Sentences out of the travel domain or containing very special meaning were removed.

6. Japanese Speech Recognition

6.1. Acoustic Model

The ML-SSS algorithm and the MDL-SSS algorithm were compared to create Japanese context-dependent HMMs. As they are described in the previous sections, they can create both contextual and temporal variations while the decision tree clustering can create contextual variations only. Furthermore, the MDL-SSS algorithm can estimate almost the best model automatically. On the other hand, the ML-SSS algorithm needs to find the best parameters by some experiments.

For the acoustic training set, we used the Japanese travel dialogs in “The Travel Arrangement Task (TRA)” of the ATR spontaneous speech database [9]. This corpus consists of role-playing pseudo-dialogs between a hotel clerk and a customer about room reservations, cancellation, trouble-shooting, etc. We also used 503 phonetically balanced sentences (BLA) read by the same 407 speakers of the TRA. TRA includes about 5 hours of speech and BLA includes about 25 hours of speech. We do not have actual in-domain speech database for the Japanese BTEC corpus. However, the TRA corpus includes a lot of similar expressions to the BTEC, and the BLA corpus is helpful to create Japanese standard phoneme models.

For analysis conditions, the frame length was 20 ms and the frame shift was 10 ms. 12 order MFCC, 12 order ΔMFCC, and Δ log power were used as feature parameters. The cepstrum mean subtraction was applied to each utterance. We used 26 kinds of Japanese phonemes and one silence. Table 1 shows the phoneme units for our Japanese ASR. A silence model with three states was built separately from the phoneme models. Three states were used as the initial model for each phoneme. The scaling factors $C_c = 2, C_t = 20$ were used for the MDL-SSS. After a topology was obtained by each topology training method, mixture components were increased, and a five Gaussian mixture model was created.

Table 1: Phoneme units for Japanese ASR

Vowels	a,i,u,e,o
Consonants	b, ch, d, g, f, h, j, k, m, n, ng, p, q, r, s, sh, t, ts, w, z, zh

6.2. Language Model and Decoding

A part of the BTEC training database consisting of 160k sentences including 1.2M words was selected as language model training data. The Multi-Class Composite 2-gram model with 1700 classes for each direction and the word 3-gram model were made from this training data. The size of the lexicon was 37k words, and the number of extracted composite words was 10k words. For recognition, the gender-dependent acoustic model and the Multi-Class bigram model were used in the first pass, and the word trigram model was used to rescore word lattices in the second pass.

6.3. Performance

For the test set, the Japanese test set 01 of the BTEC was used. It has 510 sentences. As a speech database, we collected utterances by 20 males and 20 females. Each speaker uttered 102 utterances included in the test set 01. The total number of words included in this evaluation set were 28,024.

Table 2 shows perplexity by each model. The Multi-Class Composite 2-gram obtained the middle performance between the word 2-gram model and the word 3-gram model.

Table 2: Perplexity for Japanese BTEC test set 01

word 2-gram	word 3-gram	MCC 2-gram
32.6	20.5	26.3

Table 3 shows the recognition performance represented by word accuracy rates for Japanese BTEC test set 01. We compared the models created by the ML-SSS with the maximum temporal length of states, $N_p = 3$, or 4, and the models created by the MDL-SSS. Since the initial model has three states, the models limited by $N_p = 3$ were created only by contextual splitting. Therefore, these models are considered as baseline models. There were only slightly differences of performance among these models because language models strongly fit on this evaluation data. However, both the MDL-SSS and the ML-SSS with $N_p = 4$ obtained better performance than the baseline models, the ML-SSS with $N_p = 3$. Although we concluded that the MDL-SSS can obtain almost the best model by using scaling factors, $C_c = 2$ and $C_t = 20$, it was difficult to estimate the optimal model by the MDL-SSS algorithm because the acoustic training data was out-domain data. However, the model created by the MDL-SSS with $C_c = 2$ and $C_t = 20$ obtained very close performance to the best performance by the ML-SSS.

Table 3: Recognition performance for Japanese BTEC test set 01

Topology	#states	WA [%]
ML-SSS (Np=3)	1,000	92.36
	1,400	92.48
	1,800	92.45
	2,100	92.34
	2,500	92.41
ML-SSS (Np=4)	1,000	92.60
	1,400	92.68
	1,800	92.58
	2,100	92.51
	2,500	92.36
MDL-SSS	1,132 ($C_c = 6, C_t = 20$)	92.62
	1,417 ($C_c = 4, C_t = 20$)	92.63
	1,690 ($C_c = 3, C_t = 20$)	92.55
	2,086 ($C_c = 2, C_t = 20$)	92.56
	3,193 ($C_c = 1, C_t = 20$)	92.38

7. English Speech Recognition

7.1. Acoustic model

In contrast to the Japanese language system, in-domain acoustic training data were not available to us at this time. However, as Lefevre et.al. demonstrated in [11], out-of-domain speech training data do not cause significant degradation of the system performance. It was found to be more sensitive to the language model domain mismatch. Thus, we choose the Wall Street Journal (WSJ) corpus, since we needed a speech database that is large enough and contains clean speech from many speakers. About 37500 utterances recommended for speaker-independent training (WSJ-284) were selected as the training set for our acoustic model. The total number of speakers is 284 (143 male and 141 female). Feature extraction parameters were the same as for the Japanese language system: 25 dimensional vectors (12 MFCC + 12 Delta MFCC + Delta pow) extracted from 20 ms. long windows with 10 ms. shift. First, we trained a model of the same size and topology training method as the Japanese baseline, i. e. 1400 states with 5 mixture components per state and ML-SSS algorithm. It was rather small compared to the other models that have been built on the same data [12], so it was not expected to have high performance. Nevertheless, we regarded it as a starting point for the further model development and optimization. Next, we trained several models using MDL-SSS algorithm where the temporal splitting constant C_t is set to 20 and contextual splitting constant C_c takes values from 2 to 10. In this way, we obtained models with state number ranging from about 1500 to about 7000. Initially, they all had 5 mixture components per state. The preliminary tests showed that the model with 2009 states was the best and therefore it was selected for the further experiments. Two more versions of this model: with 10 and 15 mixture components per state, were trained as well.

7.2. Language model

The language model training data consist of 600 000 sentences and about 3.4M words. Standard bigram and trigram models were trained as well as one Multi-class composite word bigram model. the number of classes is 8000 while the number of composite words is about 4000.

7.3. Pronunciation dictionary

Although the BTEC task domain is quite broad, there are many travel oriented words that are not included in the publicly available pronunciation dictionaries. Also, there are lot of specific proper names of sightseeing places, restaurants, travel related companies and brand names. A big part of the task word list represents Japanese words including Japanese first and family names. In total, there were about 2500 such words (≈10% of the 27K dictionary) and to develop good pronunciations variants for them was quite of a challenge for us. Especially difficult were the Japanese words because there is no any principled way to predict how would a native English speaker pronouns given Japanese word. This will highly depend on the speaker's Japanese proficiency with the two extremes of being fluent in Japanese and speaking just couple of wide-known words. So, we decided to cover at least these two cases by taking one pronunciation from the Japanese dictionary and converting it to the English phone labels, and generating one pronunciation according to the English phonetic rules. The later was done by using TTS software ”Festival” followed by a manual correction of some of the pronunciations judged as ”making no sense”.

Our phoneme set consists of 44 phonemes including silence. They are the same as those used in the WSJ corpus official evaluations because in this way we could use its dictionary as a source of pronunciations base-forms and also, we could run the WSJ task tests with our model for performance comparison.

7.4. Performance

In the first series of experiments, we evaluated the performance of the several acoustic models we have trained. The test data consisted of 1200 sentences from 35 speakers. Small conventional bigram and trigram language models covering about 25% of whole text training data were used in order to speed up the evaluation. The recognition results in terms of word accuracy are given in Table 4. As can be seen, the MDL-SSS model with 2009 states and 15 mixture components was the best one and it was used for the next experiments involving different types of language models.

Next, we evaluated the language models perfor-

Table 4: Acoustic models performance comparison.

Model	ML-SSS	MDL-SSS			
State #	1400	1578	2009		3028
Mix. #	5	5	5	15	5
Acc.(%)	87.5	88.1	88.5	89.4	88.2

mance. In these experiments, we used 204 utterances taken randomly from the larger BTEC test set. The results are summarized in Table 5.

Table 5: Language models performance comparison.

Model	Word 2-gram		MCC 2-gram	
3-gram Rescore	No	Yes	No	Yes
Acc.(%)	89.21	92.35	89.63	93.29

8. Chinese Speech Recognition

8.1. Acoustic Model

The basic sub-word units for the Chinese speech recognition front-end used are the traditional 21 Initials and 37 Finals:

Unit	Types
Initials	b,p,m,f,d,t,n,l,g,k,h,j,q,x,z, c,s,zh,ch,sh,r
Finals	a,ai,an,ang,ao,e,ei,en,eng,er,i1,i2,i3,ia,ian iang,iao,ie,ing,in,iu,iong,o,ou,u,ua,uai,uang uan,ui,un,uo,ong,v,van,ve,vn

The acoustic model was developed using a well-designed speech database: ATR Putonghua (ATRPTH) speech database of 2003 [13]. The database has a rich coverage of triplet Initial/Finals phonetic context, and sufficient samples for each triplet with respect to balanced speaker factors including gender and age.

The phonetically rich sentence set of ATRPTH has 792 sentences. An investigation on the *token coverage rates* have been carried out on one month's daily newspaper for different kinds of phonetic units. Table 6 shows the results, where

- Unit *A*: stands for the tonal syllable.
- Unit *B*: for the base syllable without tone discrimination.
- Unit *C*: for the normal Initial/Final triplets.
- Unit *D*: for the context tying Initial/Final triplets, which are tied based on the phonetically articulatory configurations. They are assumed to cover the major variants of triplet phonetic context[14].

Table 6: Token coverage rates of different subword units.

Unit	792 set	Newspaper	Token Coverage
A	974	1,306	98.81%
B	402	408	99.99%
C	10,906	48,392	70.15%
D	4,653	4,598	99.42%

The speakers were designed to have a balanced coverage of different genders and ages. Each unique triplets have at least 46 tokens in the speech database, guaranteeing a sufficient estimation for each triplet HMM.

During the model estimation, accurate pause segmentation and context dependent modeling [15] have been done iteratively to guarantee the model accuracy and robustness. The HMnet structure was derived through phonetic decision tree based maximum likelihood state splitting algorithm. The acoustic feature vector consists of 25 dimensions: 12 dimensional normal MFCCs, their first order deltas and the delta of frame power. The baseline gender dependent HMnets have 1,200 states, with 5 Gaussian mixtures at each state.

8.2. Language Model

The language model for Chinese ASR also uses the composite Multi-classes N-gram model. The basic lexicon has 19,191 words. The text BTEC Chinese corpus has 200 thousands of sentences for LM training. After they are segmented and POS tagged, word clustering was investigated based on the right and left context Markov dependencies. A normal word based bi-gram model showed 38.4 perplexity for the test set with 1,500 sentences. With a clustering of 12,000 of word classes, the composite multi-class bigram model showed 34.8 perplexity for the same test data. The bigram language model was used to generate word lattice in the first pass, and a trigram language model with a perplexity of 15.7 was used to rescore the word lattice.

8.3. Performance

The evaluation data is the BTEC Chinese language-parallel test data. It has 11.59 hours of speech by 20 females and 20 males. The ages of the speakers range from 18 to 55 years old. All the speakers spoke Chinese Putonghua, with little dialect accent. Table 7 shows the gender dependent, Chinese character based recognition performances. The total performance is 95.1% Chinese character accuracy. The urgent work in the near future is to fasten the searching speed without harming the recognition performance.

9. Conclusion

In this paper, we described the multi-lingual large vocabulary continuous speech recognition engine which is a part of the speech-to-speech translation system

Table 7: Chinese character based recognition performance.

Group	Character Corr.	Character Acc.
Male	96.1%	95.7%
Female	95.2%	94.4%
Total	95.7%	95.1%

being developed at ATR. Using state-of-the-art techniques like MDL-SSS acoustic model training and Multi-dimensional class N-gram language modeling we were able to build high-performance speech recognition system. However, it has been evaluated under quite clean conditions that are far different from the real world environments where the speech signal is often contaminated with noise and distorted by communication channels and room reverberation. Our current and feature work is directed to address the negative impact of such conditions on the system performance.

10. Acknowledgments

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled ”A study of speech dialog translation technology based on a large corpus”.

11. References

[1] S. J. Young, J. J. Odell, and P. C. Woodland, “Tree-based state tying for high accuracy acoustic modeling,” Proc. of the ARPA Workshop on Human Language Technology, pp. 307–312, 1994.

[2] J. Takami, S. Sagayama, “A Successive State Splitting Algorithm for Efficient Allophone Modeling,” Proc. ICASSP’92, vol. 1, pp. 573–576, 1992.

[3] M. Ostendorf, H. Singer, “HMM topology design using maximum likelihood successive state splitting,” Computer Speech and Language, vol. 11, pp. 17–41, 1997.

[4] S. Ikeda, “Construction of Phone HMM Using Model Search Method,” IEICE, vol. J78-D-II, no. 1, pp. 10–18, 1995. (In Japanese)

[5] Koichi Shinoda, Takao Watanabe, “MDL-based context-dependent subword modeling for speech recognition,” The Journal of the Acoustical Society of Japan (E), vol. 21, no. 2, pp. 79–86, 2000.

[6] S. S. Chen, R. A. Gopinath, “Model Selection in Acoustic Modeling,” Proc. of EUROSPEECH’99, vol. 3, pp. 1087–1090, 1999.

[7] Wu Chou, W. Reichl, “Decision Tree State Tying Based on Penalized Bayesian Information Criterion,” Proc. of ICASSP’99, vol. I, pp. 345–348, 1999.

[8] T. Jitsuhiro, T. Matsui, S. Nakamura, “Automatic Generation of Non-Uniform HMM Topologies Based on the MDL Criterion,” IEICE Trans. Inf. & Syst., vol. E87-D, no. 8, 2004.

[9] T. Takezawa, T. Morimoto, Y. Sagisaka, “Speech and Language Databases for Speech Translation Research in ATR,” Proc. of the 1st International Workshop on East-Asian Language Resources and Evaluation (EALREW’98), 1998.

[10] H. Yamamoto, Y. Sagisaka, “Multi-Class Composite N-gram Based on Connection Direction,” Proc. of ICASSP’99, vol. 1, pp. 533–536, 1999.

[11] F. Lefevre, J.L. Gauvain and L. Lamel, ”Improving Genericity for Task-independent Speech Recognition”, in *Proc. EuroSpeech*, pp.1241-1244, 2001.

[12] SPOKEN LANGUAGE TECHNOLOGY, *Proc. of a Workshop*, Plainsboro, New Jersey, March 6-8, 1994.

[13] J.-S. Zhang, M. Mizumachi, F. K. Soong and S. Nakamura, ”An introduction to ATRPTH: a phonetically rich sentence set based Chinese Putonghua speech database developed by ATR”, Proc. of the he Acoustic Society of Japan (ASJ), pp.167-168, 2003 Fall.

[14] J.-S. Zhang, Sh.-W. Zhang, Y. Sagisaka, and S. Nakamura, ”A hybrid approach to enhance task portability of acoustic models in Chinese speech recognition”, Proc. of Eurospeech2001, Denmark, Sept. 2001. Vol.3, pp.1661-1663.

[15] J.-S. Zhang, K. Markov, T. Matsui and S. Nakamura, ”A study on acoustic modeling of pauses for recognizing noisy conversational speech”, IEICE Transaction on INFO. and SYST., Vol.86-D, No. 3, March 2003.