Computer Speech and Language (1999) **13**, 79–97
Article No. csla.1998.0112
Available online at http://www.idealibrary.com on IDEAL

Automatic selection of phonetically distributed sentence sets for speaker adaptation with application to large vocabulary Mandarin speech recognition

Jia-lin Shen,[†] Hsin-min Wang,[†] Ren-yuan Lyu[‡][§] and Lin-shan Lee[†][‡][¶]

[†]*Institute of Information Science, Academia Sinica,* [‡]*Dept of Electrical Engineering, Chang Gung College of Medicine and Technology, Taipei, Taiwan, Republic of China,* [§]*Dept of Electrical Engineering, National Taiwan University*

Abstract

This paper presents an approach of automatic selection of phonetically distributed sentence sets for speaker adaptation, and applies the concept to the task of Mandarin speech recognition with very large vocabulary. This is a different approach to the adaptation data selection problem. A computer algorithm is developed to select minimum sets of phonetically distributed training sentences from a text corpus defining the desired task. These sentence sets not only include an almost minimum number of words and sentences that cover the desired acoustic units, but also have statistical distributions of these acoustic phonetic units very close to that in the given text corpus defining the desired task. In this way, more frequently used units can be better trained with higher accuracy, thus improving the overall performance, but the new user needs to produce only a small number of meaningful sentences to train the recognizer. Different sets of sentences selected using different phonetic criteria taking into consideration the statistics of the different acoustic units in the given corpus can then be integrated into a multi-stage adaptation procedure. With this procedure, the recognition performance can be improved incrementally stage by stage using the adaptation data produced with these sentence sets. This proposed approach is applied to an example task of Mandarin speech recognition with a very large vocabulary, both in isolated syllable and continuous speech modes and includes different subject domains in continuous speech recognition. Although the primary results obtained in this paper are for this example task, it is believed that many of the concepts and techniques developed here will also be very useful for other speaker adaptation problems and other languages.

[¶]Author for correspondence.

0885–2308/99/010079 + 19 $30.00/0

1. Introduction

The adaptation of a speech recognizer to the signal characteristics of a new user using the minimum amount of training data produced by the new user has been a very important issue for speech recognition for a long time. In general, speaker adaptation techniques can be categorized into two classes: feature-based (Shikano, Lee & Reddy, 1986; Stern & Lasry, 1987; Furui, 1989; Huang, 1992; Matsukoto & Inoue, 1992) and model-based (Lee, Lin & Juang, 1991; Huang & Lee, 1993; Gauvain & Lee, 1994; Hao & Fang, 1994; Zhao, 1994; Leggetter & Woodland, 1995) approaches. The former tries to modify the feature vectors of the input speech of the new user, while the latter tries to modify the model parameters used for recognition instead of the features. Typical examples for the latter type of model parameter adaptation include direct estimation of model parameters by maximum *a posteriori* (MAP) algorithms (Lee *et al.*, 1991; Gauvain & Lee, 1994) as well as indirect transformation-based estimation of model parameters by maximum likelihood linear regression (MLLR) algorithms (Leggetter & Woodland, 1995). For such model-based approaches, there are two other issues to tackle: the seed model and the adaptation data selection. For seed model selection, gender-dependent speaker-independent models are usually used as the initial model for the refinement of the model parameters. Speaker clustering techniques are also useful for the classification of different speakers so that better initial models can be obtained for a new speaker (Kosaka *et al.*, 1996). For adaptation data selection, words or sentences rich in phonetic information should be chosen for the new user to produce the adaptation data so that a significant improvement can be obtained using a minimum amount of adaptation data (Yu & Liu, 1990; Jan, van Santen & Buchsbaum, 1997). In this way, the adaptation data may become task-dependent, i.e. it could be dynamically selected for a specific task. Adaptation data selection becomes critical for tasks with a very large vocabulary, because very often only a limited improvement can be achieved with reasonable amount of adaptation data for such tasks due to the complexity of the model. Some techniques have been developed to overcome this problem. For example, predictive speaker adaptation techniques (Cox & Bridle, 1995) such as vector-field smoothing (VFS) (Hattori & Sagayama, 1992; Takahashi & Sagayama, 1997) and adaptive Bayesian learning algorithms (Huo & Lee, 1997) have been proposed for the improvement of those models without corresponding adaptation data. Also, the transformation-based adaptation algorithms (Leggetter & Woodland, 1995; Digalakis & Neumeyer, 1995) were developed such that different models can share the same transformation matrix obtained for those models with adaptation data. These methods can be further integrated to achieve better performance.

In this paper, a different approach to the adaptation data selection problem is considered. A computer algorithm is developed to select minimum sets of phonetically distributed training sentences from a text corpus defining the desired task. These sentence sets not only include almost the minimum number of words and sentences that cover the desired acoustic units, but also have a statistical distribution of these desired acoustic units very close to that in the given text corpus defining the desired task. In this way, more frequently used units can be better trained with higher accuracy, thus improving the overall performance, but the new user needs to produce only the smallest number of meaningful sentences to train the recognizer. Different sets of such sentences selected based on different phonetic criteria considering the statistics of different acoustic units in the given corpus can then be integrated into a multistage adaptation procedure. With this procedure, the recognition performance can be improved incrementally stage-by-stage using the adaptation data produced with these sentence sets. This proposed approach is applied to an example task of Mandarin speech recognition with very large vocabulary, both in isolated syllable and continuous speech modes, including different subject domains in continuous speech recognition. Although the primary results obtained in this paper are for this example task, it is believed that many of the concepts and techniques developed here will also be very useful to other speaker adaptation problems and other languages.

Mandarin Chinese is a monosyllablic tonal language. There are at least 100,000 commonly used words and each word is composed from one or several characters. There are also at least 10,000 commonly used characters. However, all the Chinese characters are pronounced as monosyllables, of which there are a total of only 1345 different phonologically allowed syllables. This monosyllabic structure is a characteristic feature of Mandarin Chinese when recognition of very large vocabulary is considered. Moreover, every syllable is assigned a tone, and the tone has lexical meaning. When the differences in tone are disregarded, these 1345 different "tonal syllables" are further reduced to 408 different "base syllables" (i.e. tone-independent syllable structures). Note that here we use "tonal syllables" to indicate syllables with tones, but "base syllables" for those disregarding the tones. Because there are only four lexical tones plus a neutral tone and the tones can be independently recognized using primarily pitch information, accurate recognition of all the 408 Mandarin base syllables is believed to be the key problem in Mandarin speech recognition with very large vocabulary. This is the example task selected in this paper, i.e. the recognition of the 408 base syllables in Mandarin speech, to demonstrate the concepts and techniques proposed here. When the base syllables information is combined with the tone information, the corresponding Chinese characters and therefore words and sentences can be obtained by a lexical access process and a Chinese language model (Lee *et al.*, 1993*a*, *b*). Because the purpose of the example task is simply to recognize the base syllables, the tones were not considered and no language model was used in the example experiments described below.

This paper is organized into eight sections. In Section 2, the computer algorithm to select phonetically distributed sentence sets to produce the adaptation data is developed. Typical sentence sets are selected for the example task of Mandarin Chinese in Section 3. In Section 4, the simplified on-line adaptation algorithm used in the following experiments is presented. The speech database used in the experiments is described in Section 5, and the experimental results for isolated syllable and continuous speech recognition are then discussed in Sections 6 and 7, respectively. The concluding remarks are given in Section 8.

2. Automatic selection of phonetically distributed sentences for adaptation data

In this section, we present an algorithm to select automatically the smallest set of sentences from a given text corpus for a given task. Certainly these sentences have to include all desired acoustic units for the given task, and therefore they are phonetically balanced. Furthermore, all the acoustic units should appear in this set of sentences with a statistical distribution approximating the distribution of these units in the given corpus for the given task. In this way the more frequently used acoustic units will be trained better and recognized more accurately, and thus the overall recognition accuracy for this given task can be improved, or the desired accuracy can be achieved faster. Such a sentence set sometimes can also be used as a good testing set for evaluation of the performance of a recognizer with respect to a specific given recognition task. Because of the similar statistical distribution for the acoustic units, the test results for this set of sentences may give a better reflection of the real performance of a recognizer for a given task. This specific statistical distribution of the acoustic units for a given task is a specific feature of the "phonetically distributed" sentence set mentioned here. With the rapid development of speech recognition technology today, the variety of speech recognition tasks (applications, subject domains, vocabularies, sentence patterns, etc.) are growing very

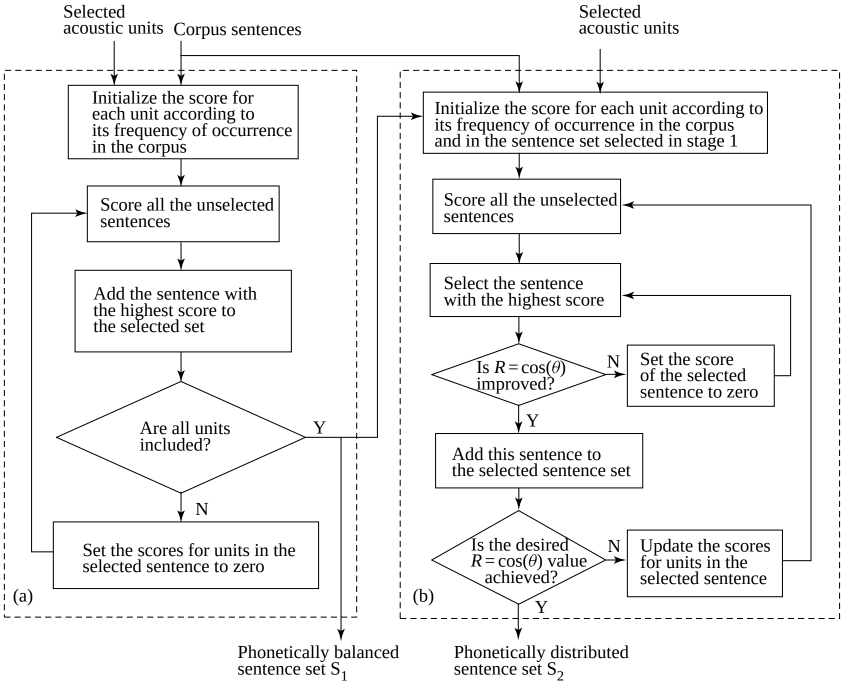
Selected acoustic units
Corpus sentences
Selected acoustic units
Initialize the score for each unit according to its frequency of occurrence in the corpus
Score all the unselected sentences
Add the sentence with the highest score to the selected set
Are all units included?
Y
N
Set the scores for units in the selected sentence to zero
(a)
Initialize the score for each unit according to its frequency of occurrence in the corpus and in the sentence set selected in stage 1
Score all the unselected sentences
Select the sentence with the highest score
Is $R = \cos(\theta)$ improved?
N
Set the score of the selected sentence to zero
Y
Add this sentence to the selected sentence set
Is the desired $R = \cos(\theta)$ value achieved?
N
Update the scores for units in the selected sentence
Y
(b)
Phonetically balanced sentence set S_1
Phonetically distributed sentence set S_2

Figure 1. The flow chart of the automatic sentence selection algorithm: (a) Phase 1; (b) Phase 2.

rapidly, and generating for each task, especially those with relatively large vocabulary, such a set of phonetically distributed sentences for speaker adaptation automatically using a computer algorithm from a given text corpus is highly desired.

An algorithm with the above described purpose is presented here. The block diagram of such an algorithm is shown in Figure 1. This algorithm has two phases. The first phase is to include all acoustic units needed for the given task with minimum number of sentences, while the second phase is to try to reproduce the statistical distribution of these acoustic units in the given text corpus for the given task with also a minimum number of sentences, i.e. to include those frequently used acoustic units more times in the sentence set and so on. This algorithm can be applied to any language and any given recognition task for any set of selected acoustic units, as long as the text corpus defining the desired recognition task is given.

The first phase of the algorithm is shown in Figure 1(a). The input is the whole text corpus defining the desired task as well as a set of selected acoustic units, while the output is a "phonetically balanced" sentence set S_1, which is almost the smallest set of sentences including all necessary acoustic units, but not necessarily with the desired statistical distribution for these units. The basic principles in this phase are the following two rules to ensure that the total number of sentences covering all acoustic units can be as small as possible:

(1) Those sentences including more distinct acoustic units should be selected with higher priority.
(2) Those sentences consisting of acoustic units with lower frequency of occurrence in the corpus should also be selected with higher priority.

To realize these basic principles, as can be seen in Figure 1(a), a score is first assigned to each selected acoustic unit, which is initialized inversely proportionally to the frequency of occurrence of the unit in the given text corpus, such that the rarely used units have higher priority to be selected. A score is then defined for each sentence in the given text corpus as well, which is primarily the average of the scores of its component units, but modified by a parameter which is higher for sentences with larger number of distinct acoustic units, because such sentences should be selected with higher priority. The sentence with the highest score is then selected and included in the desired sentence set S_1. Once a sentence is selected, the scores of all its component units are immediately set to zero to avoid these units to be selected again. The algorithm thus recursively updates the scores of the acoustic units and of all the remaining unselected sentences in the given text corpus, and selects additional sentences with the highest scores, until all the desired acoustic units are included in the selected sentence set. In this way, a set of almost minimum number of sentences S_1 which includes all acoustic units can be obtained.

The second phase of the algorithm is shown in Figure 1(b). The input for this phase includes the unselected sentences left in the original given text corpus, and the "phonetically balanced" sentence set S_1 obtained in the first phase, while the desired output is an almost smallest set of "phonetically distributed" sentence set S_2. This set not only includes all necessary acoustic units, but the statistical distribution of these units should approximate that in the given corpus for the desired task. In this phase, the score of each acoustic unit is re-defined in a different way. The purpose is that the acoustic units with higher frequency of occurrence in the original corpus and lower frequency of occurrence in the sentence set S_1 previously obtained in the first phase should have higher priority to be selected. This is the way to reproduce the statistical distribution of the acoustic units with minimum number of sentences. In order to achieve this purpose, a score reduction parameter is first defined for each acoustic unit to reduce the priority of that unit to be selected again after each time it is selected and included in the sentence set. Also, this parameter is inversely proportional to the frequency of occurrence of this unit in the original text corpus, so that frequently used units retain higher priority to be selected even if they have been selected before. The initial score for each acoustic unit in the second phase is then defined as a constant subtracted by this score reduction parameter multiplied by the times it has been selected and included in the sentence set S_1 obtained in the first phase. The rest of the algorithm is very similar to that of the first phase. However, in this phase a similarity measure R is defined to estimate the degree to which the statistical distribution of the acoustic units in the selected sentence set is similar to that in the original corpus,

$$R = \frac{\vec{v}_c \cdot \vec{v}_d}{|\vec{v}_c||\vec{v}_d|} = \cos(\theta) \tag{1}$$

where $\vec{v}_c = [n_c(1), n_c(2), \ldots, n_c(i), \ldots, n_c(L)]$, $\vec{v}_d = [n_d(1), n_d(2), \ldots, n_d(i), \ldots, n_d(L)]$, $n_c(i)$ and $n_d(i)$ are the times the i-th acoustic unit appears in the corpus and in the currently selected sentence set respectively, and L is the total number of different acoustic units. Apparently, $\vec{v}_c$, $\vec{v}_d$ represent the statistical distribution of the acoustic units in the corpus and in the currently selected sentence set respectively, R is the normalized inner product of $\vec{v}_c$ and $\vec{v}_b$, and θ is the angle between $\vec{v}_c$ and $\vec{v}_d$. When $R = 1$, i.e. $\vec{v}_c = \mathrm{k}\vec{v}_d$, the two statistical distributions will be exactly identical. Now, the unselected sentence in the text corpus with the highest score and at the same time can improve the similarity measure R is first selected and added to the sentence set. Once a sentence is selected, the scores for all its component acoustic units are immediately subtracted by their score reduction parameter multiplied by the times these units appear in the selected sentence. By recursively selecting additional sentences one-by-one as described above, a minimum set of phonetically distributed sentences with desired statistical distribution for the acoustic units can thus be obtained.

3. Incremental phonetically distributed sentence sets for speaker adaptation in Mandarin syllable recognition

Here we applied the sentence set selection algorithm mentioned above to the Mandarin syllable recognition problem discussed previously. Although the above algorithm seems simple, there exists a variety of different ways to apply it to achieve different purposes in speaker adaptation. The results below show some typical examples of such variety. There certainly exist other different approaches with different results. First, it is desirable to have incremental sentence sets such that the new user can observe some performance improvement stage-by-stage each time the utterances for an incremental sentence set are produced. Second, each incremental sentence set may be obtained via different acoustic requirements such that different purposes for speaker adaptation can be achieved. These will be demonstrated in the following examples.

Here two collections of incremental phonetically distributed sentence sets obtained using the sentence selection algorithm are presented. They are for speaker adaptation on the task of the recognition of all the 408 Mandarin base syllables in isolated syllable and continuous speech mode, respectively. Conventionally, each Mandarin syllable is decomposed into an INITIAL/FINAL format similar to the Consonant/Vowel format in other languages, where INITIAL means the initial consonant of the syllable and FINAL means the vowel (or diphthong) part but including optional medial and nasal ending. There are a total of 22 different context-independent (CI) INITIALs and 41 different context-independent (CI) FINALs for the 408 Mandarin base syllables. Due to the mono-syllabic structure of Mandarin Chinese, the inter-syllable context dependency is observed to be much less significant than intra-syllable context dependency even in continuous speech. This is why in some situations, it is desirable to consider only the latter but ignore the former to simplify the problem. In such an approach, the 22 different context-independent (CI) INITIALs can be expanded to 113 context-dependent (CD) INITIALs considering the beginning phoneme of the following FINALs (Lee *et al.*, 1993*a*). In this way, the smaller number of smaller acoustic units such as these 113 CD INITIALs and 41 CI FINALs can be used in the selection of earlier training sentence sets such that the recognition rates can be improved very quickly in earlier stages, while the larger number of larger acoustic units such as base syllables or tonal syllables can be used in the selection of later training sentence sets to train the model parameters further in later stages. This is the basic concept of using incremental training sentence sets selected with different acoustic requirements for speaker adaptation. The text corpus used here for selection of training sentences consists of a total of 124 845 sentences (1 374 182 characters) collected from daily Chinese newspapers, Chinese magazine articles and so on, covering almost all important subject areas and defining a task of general domain, although any other given corpus defining a specific recognition task is certainly equally applicable.

3.1. Adaptation sentence sets for isolated syllable recognition

The first collection of sentence sets is for isolated syllable recognition. For this purpose, a total of four phonetically distributed sentence sets, sentence sets 1, 2, 3 and 4 were selected. The first set, sentence set 1, covers all the 113 CD INITIALs and 41 CI FINALs with statistical distribution approximating to that in the corpus. This set consists of only 24 sentences or 188 syllables (characters). Apparently, 113 syllables out of the 188 cover all the 113 CD

TABLE I. The incremental sets of phonetically distributed sentences selected for (a) isolated syllable recognition and (b) continuous speech recognition in the example task

(a) Isolated syllable recognition

Set	Selected sentences		Included syllables		Acoustic units included	Coverage of syllables in given corpus
	Number	Total number from set 1	Number	Total number from set 1		
1	24	24	188	188	113 CD INITIALs/41CI FINALs	—
2	15	39	104	292	Top 100 tonal syllables	48.82%
3	22	61	144	436	Top 200 tonal syllables	69.24%
4	86	147	556	992	Top 500 tonal syllables	93.83%

(b) Continuous speech recognition

Set	Selected sentences		Included syllables		Acoustic units included	Coverage of syllables in given corpus
	Number	Total number from set 1	Number	Total number from set 1		
1	24	24	188	188	113 CD INITIALs/41CI FINALs	—
2	76	100	617	805	408 base syllables Top 20 inter-syllable Context dependency classes	—
3	100	200	801	1606	Top 600 tonal syllables Top 40 inter-syllable Context dependency classes	96.34%

INITIALs, and the additional 75 syllables take care of the statistical distribution and make up all the sentences. Similarly for the 41 CI FINALs. The second phonetically distributed sentences set, sentence set 2, however, is constructed by including 15 extra sentences or 104 extra syllables (characters) to be added to the previous sentence set 1 to form a set of 39 sentences or 292 syllables (characters). This combination of sets 1 and 2 covers the top 100 most frequently used tonal syllables out of the total of 1345 tonal syllables. In the sentence sets 3 and 4, additional 22 and 86 sentences or 144 and 556 syllables (characters) are further added, such that when combined with sets 1 and 2, they can cover the top 200 and 500 most frequently used tonal syllables out of the total of 1345, similar to that of set 2. In fact, up to the sentence sets 2, 3 and 4, the top 100, 200 and 500 most frequently used syllables out of 1345 are well trained, and they in fact already cover 48.82%, 69.24% and 93.83% of the syllables in the text corpus defining the task of Mandarin syllable recognition with general domain. All these data are summarized in Table I(a) and Figure 2(a).

3.2. Adaptation sentence sets for continuous speech recognition

In the continuous speech recognition experiments below, the 113 CD INITIALs and 41 CI FINALs are also used as the basic recognition units. The inter-syllable context dependency is not explicitly covered in the model configurations, but it is certainly desirable that they be included in the adaptation utterances and reasonably reflected in the adapted models. For this purpose the acoustic criteria used to select the incremental adaptation sentence sets are different from those used for isolated syllable recognition. All the FINALs of Mandarin syllables can be divided into 12 groups based on their ending phonemes, and all the INITIALs into seven groups based on the state of articulation. So the inter-syllable context dependency can be roughly categorized into 84 classes. Three sets of phonetically distributed sentences were thus developed for continuous speech recognition, sets 1, 2 and 3. Set 1 is exactly the same as that for isolated syllable recognition presented in Section 3.1, with 24 sentences or 188 syllables covering all the 113 CD INITIALs and 41 CI FINALs. However, in sentence set 2, the criterion is to cover all of the 408 base syllables, as well as the top 20 most frequently

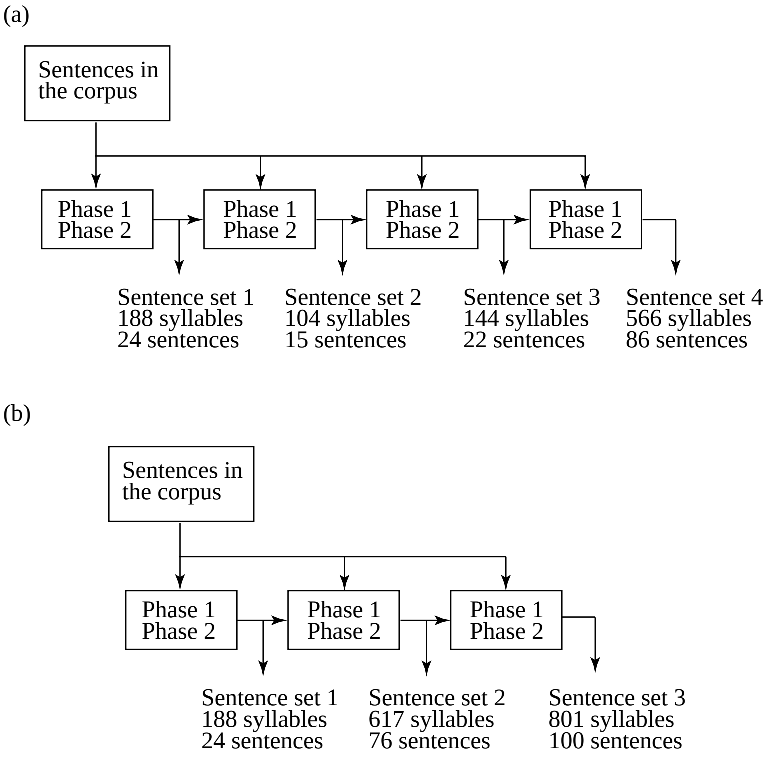

Figure 2. The block diagram of the sentence selection algorithm with (a) four stages for isolated syllable recognition; and (b) three stages for continuous speech recognition.

occurring classes of inter-syllable context dependency out of the total of 84, including the statistical distribution of these base syllables and inter-syllable context dependency classes in the given corpus. As shown in Figure 2(b), an additional 76 sentences or 622 syllables (characters) are included in the set 2. In the third stage, an additional 100 sentences or 801 syllables (characters) are added to cover the top 600 tonal syllables as well as the top 40 classes of inter-syllable context dependency, with statistical distributions approximating to those in the corpus. In this way, 96.34% of syllables and 92.51% of inter-syllable context dependency classes in the given corpus are included with a total of 200 sentences or 1606 syllables (characters). Table I(b) summarizes the statistical data of this collection of three phonetically distributed sentence sets for continuous speech recognition.

Note that these automatically selected sentence sets are quite different from the randomly selected sentences usually used in many adaptation tasks. With the proposed algorithm here, the selected sentence sets cover all desired acoustic units with given distribution using only a minimum number of words or sentences based on given acoustic criteria. So the most frequently used acoustic units and events can have adaptation utterances available as early as possible and repeated for as many times as needed and thus can be better trained and recognized more accurately in the earliest stage. This is the way to generate most efficient adaptation data. In a set of randomly selected sentences, on the other hand, some acoustic units may be missing, some important frequently used units may appear only very seldom, resulting in under-training, while some less frequently used units may appear repeatedly which is a waste of the target user's time. Of course the correct statistical distribution of the units should automatically appear in randomly selected sentences as well in principle, but that becomes true only after a large number of randomly selected sentences are included, which makes the adaptation less efficient. Also, with the proposed algorithm various desired acoustic criteria can be set by the designer to achieve various purposes. For example, with the staged adaptation design, the adaptation goal may be achieved most efficiently at the earliest time by a few sub-goals stage by stage. All these will be verified later on in this paper by the example experimental results.

4. A simplified on-line adaptation algorithm

In many applications, it is an attractive and desired feature of a recognition system that the accuracy can be improved incrementally when the user produces the adaptation utterances one-by-one. In this way, the performance for this user can be continuously improved when he keeps training or using the system. A supervised on-line adaptation algorithm for this purpose is summarized below (Shen, Wang, Lyu, & Lee, 1994), which will be used in the following experiments in this paper.

If the adaptation data Y is composed of T utterances, i.e. $Y = y_1, y_2, \dots, y_T$, the adapted model $\hat{\lambda}^{(T)}$ given the last utterance y_T can be estimated as the following (Huo & Lee, 1997):

$$\hat{\lambda}^{(T)} = \arg\max_{\lambda} p(y_T|\lambda)p(\lambda|y_1, y_2, \dots, y_{T-1})\cdot \qquad (2)$$

In other words, the adapted model obtained with all the previous adaptation data given can be used as the seed model for adaptation with the next utterance. Then the adapted mean parameters can be estimated as the linear interpolation of the mean parameters for the current model and the new utterance as in the following form (Huo & Lee, 1997):

$$\eta_j^{(ad)}(t+1) = \eta_j^{(ad)}(t) + \tau(t)(\eta_j^{(n)}(t) - \eta_j^{(ad)}(t)), \qquad t = 1, \dots, T, \qquad (3)$$

where $\eta_j^{(n)}(t)$ and $\eta_j^{(ad)}(t)$ are the mean parameters for the new adaptation utterances and the adapted model respectively at time t, j is the mixture index, and $\tau(t)$ is the interpolation factor at time t. The corresponding covariance matrix can be updated accordingly. As long as $\tau(t)$ monotonically decreases with t and satisfies the following two conditions:

$$\sum_{t=1}^{\infty} \tau(t) \to \infty, \qquad \sum_{t=1}^{\infty} \tau(t)^2 < \infty, \qquad (4)$$

the adapted models can be continuously improved. This is the algorithm to be used in the following experiments, the concept coming from the learning vector quantization algorithm (Kohonen, 1988). With this algorithm, the model parameters can be continuously adapted on-line in real time each time the new speaker produces either a single utterance or a set of utterances due to the low computation complexity involved.

5. Speech database used in the experiments

The speech database used in all experiments presented below in this paper is categorized into two sets. The first data set was produced in isolated syllable model for isolated syllable recognition, while the second set was produced in continuous speech mode for continuous speech recognition. The first set in isolated syllable mode was produced by two groups of speakers with a total of 43 male speakers and 31 female speakers. The first speaker group A1 includes three male speakers. Each speaker of group A1 produced all the 1345 Mandarin tonal syllables in isolated syllable mode with four utterances for each syllable. In addition, each speaker of group A1 also produced the syllables in the first collection of four incremental sets of phonetically distributed sentences for isolated syllable recognition as mentioned in Section 3.1, also in isolated syllable mode with one utterance for each syllable. These three speakers in group A1 were used as testing speakers in the following experiments. The second speaker group A2 consists of 40 male speakers and 31 female speakers. These 71 speakers produced all the 1345 Mandarin tonal syllables each with 39 utterances, so the total number of syllable utterances obtained from speaker group A2 is $1345 \times 39 = 52\,455$, in which each individual speaker uttered about 739 syllables on average. These speech data produced by group A2 speakers were used to train a speaker-independent model. The average speaking rate for this set of speech database is 0.42 s/syllable.

On the other hand, the second data set in continuous speech mode for continuous speech recognition contains data produced by two groups of speakers with a total of 70 male speakers and 55 female speakers. The first speaker group B1 consists of three male speakers used as the testing speakers. Each of the three speakers produced six paragraphs of texts covering six subject domains of business, politics, society, philosophy, science and sports, respectively, with a total of 134 sentences or 1462 characters. These data were used for testing. In addition, each speaker of group B1 also produced the second collection of three incremental sets of phonetically distributed sentences for continuous speech recognition as described above in Section 3.2. These data were used as the adaptation data. In the second speaker group B2, a total of 14 080 continuous speech sentences were produced by the rest of 67 male speakers and 55 female speakers. These data produced by group B2 were used to train a speaker-independent model. The average speaking rate for this set of speech database is 0.25 s/syllable.

All the speech data were obtained in an office-like laboratory environment. They are low-pass filtered, digitized by an Ariel S-32C DSP board with sampling frequency 16 kHz. After end-point detection is performed, a 20 ms Hamming window is applied every 5 ms with a pre-emphasis factor of 0.95. The speech features used in the following experiments include 14 order cepstral coefficients and the corresponding 14 delta cepstral coefficients.

6. Experimental results for isolated syllable recognition

In the isolated syllable recognition experiments here, the Segmental Probability Model (SPM) is used (Lyu, Hong, Shen, Lee & Lee, 1998). This model is very similar to continuous density HMM (CHMM) with Gaussian mixtures, except that the state transition probabilities are deleted and the N states simply equally segment the syllable utterances. With the SPM, almost identical recognition accuracy can be obtained for isolated Mandarin base syllables but at significantly reduced computation complexity compared with CHMM, due to the relatively simple phonetic structure of these syllables. In the adaptation process, the segment sharing concept based on the phonological structure of Mandarin syllables is applied, which is also very similar to the tied-state method used in HMM (Lee, Giachin, Rabiner, Pieraccini, & Rosenberg, 1992). In this approach, the first few segments (or states) of the SPMs representing the INITIAL parts of the syllables having the same CD INITIALs share the same adaptation data, and so do the last few segments of the models representing the FINAL parts of the syllables having the same CI FINALs. Also, in the series of isolated syllable recognition experiments to be discussed in this section, to reflect better the realistic performance of a speech recognition system with respect to the desired task of general domain defined by

TABLE II. The adaptation results for isolated syllables using manually selected utterances with adaptation data varying from 113 to 1345 × 3 utterances

	Adaptation with number of available utterances					
	SI	113	408	1345	1345 × 2	1345 × 3
Bayesian adaptation (BA)	63.05%	79.70%	86.72%	90.94%	92.30%	94.14%
Simplified on-line	63.05%	77.52%	86.65%	90.63%	91.49%	93.88%
Simplified on-line plus BA	63.05%	79.23%	88.34%	91.59%	92.28%	95.08%

the given corpus, all the recognition results mentioned are the average of the recognition accuracies for all the individual syllables, but weighted by the frequency of appearance for the syllables evaluated from the given text corpus. So the correct recognition of a frequently used syllable will be counted more than that of a rarely used syllable.

Speaker-dependent tests for the three testing speakers in group A1 were first performed to obtain baseline results for comparisons later on. Three utterances for each of the 1345 Mandarin tonal syllables for each speaker in group A1 were used in training and the last utterance in testing. The *leave-one-out* method was applied, i.e. among the four utterances available for each syllable for each speaker, three utterances were used in training and the last one in testing. This operation was repeated four times so all the four utterances have been used as testing data. The average result for the four tests is taken as the final result for the speaker. The average for the three speakers for this speaker-dependent experiment is 92.95% (this number is 92.65% if CHMM with the same model configuration is applied instead). Speaker-independent tests were also performed to obtain another comparison baseline. The speaker-independent models were trained using all the speech data produced by all the 71 speakers in group A2, and the average recognition accuracy for the three test speakers in group A1 is found to be 63.05% (this number is 62.76% if CHMM is applied instead).

6.1. *Adaptation results for manually selected training syllables*

Apparently, for isolated syllable recognition, one does not need to use the sentence sets selected above in Section 3.1, but can use manually selected syllables for adaptation. The minimum set of manually selected training utterances is 113 syllables constructed by the 113 CD INITIALs and 41 CI FINALs selected from the most frequently used syllables. The next set is the 408 base syllables with the most frequently used tone for each base syllable. One can also use one, two or three sets of all the 1345 tonal syllables because they are available. Experiments for such manually selected training syllables were performed first for comparison. The results are summarized here.

The results based on Bayesian adaptation algorithm (Gauvain & Lee, 1994) with segment sharing are listed in the first row of Table II with the available adaptation data for each test speaker varying from 113 up to 1345 × 3 syllables. It can be found that the 113 adaptation syllable utterances can achieve a significant improvement with 45.06% error rate reduction (accuracy from 63.05% of speaker-independent case to 79.70%). A single utterance for each CD INITIAL and CI FINAL is very efficient but not adequate. When 408 base syllables are included in the adaptation data, the accuracy is improved to 86.72%. When one, two and three utterances of all the 1345 utterances are available, accuracies of 90.94%, 92.30% and 94.14% can be achieved. The last case (94.14%) is even better than the speaker-dependent case mentioned previously (92.95%).

The corresponding recognition rates using the simplified on-line adaptation algorithm presented in Section 4, in which adaptation was performed each time an additional adaptation utterance was included, are listed in the second row of Table II. In this way, the rates can actually be improved utterance-by-utterance, and the process can be performed every time a new utterance is entered to the system. In comparison with the first row, the rates for this approach are slightly lower, but the differences are gradually reduced as the number of available adaptation utterances is increased. In the last row of Table II, the adapted models with the simplified on-line adaptation algorithm mentioned above are further re-estimated by the Bayesian adaptation at the end of each stage when 113, 408, 1345 utterances have been entered and so on. As can be seen, the performance of the simplified on-line adaptation algorithm can be further improved in this way and even outperform the results obtained directly by the Bayesian adaptation in most cases.

6.2. Adaptation results for the selected phonetically distributed sentence sets

In the second set of experiments, the syllable utterances produced for the first collection of four sets of carefully selected phonetically distributed sentences presented in Section 3.1 were used for adaptation. The system can be adapted to a new speaker in four stages. In each stage, the utterances produced for one sentence set were used, and the recognition rate can be improved stage-by-stage incrementally. In fact, when the simplified on-line adaptation technique described in Section 4 is used, the recognition rate can also be improved not only stage-by-stage, but utterance-by-utterance. Table III is the experimental results for such a four-stage incremental procedure, in which in each stage the simplified on-line adaptation algorithm is first performed after each utterance was entered, with results listed as experiment 1 in the left part of the table, and the model parameters are then further re-estimated using the Bayesian adaptation after all utterances of the stage have been entered, with results listed as experiment 2 in the right-hand column of the table, evaluated for the three outside speakers in group A1 respectively. So the results in experiment 2 of Table III are somewhat similar to the data in the last row of Table II, but with different sets of adaptation utterances. It can be seen from Table III that the average initial recognition rate for the speaker-independent model is 63.05%, as in the first row, although the actual rates vary quite significantly across different speakers. After the first stage with the 188 characters or 24 sentences produced by a new speaker (with total length of speech data within 1.3 min), the second row of Table III shows that the average recognition rate is immediately improved significantly to 82.28%, and it can be further improved to 84.00% after re-estimation of the model parameters using the Bayesian adaptation. Then with the next three stages completed with the sentence sets 2, 3 and 4 including a total of 147 sentences, 992 utterances or 6.9 min of adaptation speech signal counted from the first stage, the finally achieved recognition rates can be as high as 92.96% and 93.64% in experiments 1 and 2 respectively, even slightly exceeding the speaker-dependent results (92.95%) mentioned previously trained by $1345 \times 3 = 4035$ syllable utterances. Another nice feature here is that the achieved recognition rates at each stage are in fact much more stable across different speakers, although this is not the case for the initial rates for speaker-independent models. Moreover, with model parameters re-estimated further with the same adaptation data by the Bayesian adaptation after each stage, the recognition rates in experiment 2 can be further improved by $1 \sim 2\%$ in each stage compared with those in experiment 1.

Figure 3 shows the learning curve in full lines for the four-stage adaptation procedure, i.e. rate increase as a function of the number of adaptation utterances used based on the results of experiment 2 in Table III. As a comparison, the results obtained by a similar procedure as listed in the last row of Table II using manually selected 113, 408, 1345, 1345×2 and 1345×3 adaptation utterances are also plotted in Figure 3. It is noteworthy that the learning slope of

TABLE III. The four-stage adaptation results for isolated syllable recognition using the four sets of phonetically distributed sentences

	Total length of speech signal from stage 1	Experiment 1				Experiment 2
		Speaker 1	Speaker 2	Speaker 3	AVE	AVE
SI results	—	59.95%	59.26%	69.95%	63.05%	63.05%
Stage 1	1.3 min	82.47%	81.36%	83.02%	82.28%	84.00%
Stage 2	2.0 min	89.02%	86.82%	87.58%	87.81%	88.75%
Stage 3	3.1 min	91.34%	89.00%	90.24%	90.19%	92.18%
Stage 4	6.9 min	93.75%	93.43%	91.70%	92.96%	93.64%

the four-stage phonetically distributed sentence sets is almost always higher than that using the manually selected adaptation utterances, except for the beginning 113 manually selected utterances. Obviously, the highest learning slope is obtained when the first 113 manually selected utterances covering all the 113 CD INITIALs and 41 CI FINALs are used, because this is the most compact collection of utterances covering all the necessary units. However, the first phonetically distributed sentence set of 24 sentences or 188 syllables is equally applicable to continuous speech recognition case, as this set is also used as set 1 in the phonetically distributed sentence sets for continuous speech recognition presented in Section 3.2, because they are meaningful sentences thus can be produced continuously. The 113 manually selected syllables, on the other hand, cannot be extended to continuous speech case, because they do not form meaningful sentences. In fact, the learning slope obtained by the first set of phonetically distributed sentences with 188 syllable utterances is also very high, only slightly lower than that using the 113 manually selected utterances. But this stage of 188 utterances gives a much higher recognition rate than the 113 manually selected utterances (84.00% as compared to 79.23%), apparently because these 188 utterances include the statistical distribution of the INITIAL/FINALs such that more frequently used units are trained better with higher accuracy, thus much better overall performance can be achieved. This is the apparent advantage of properly utilizing the statistics of the acoustic units in adaptation processes. In the second stage of phonetically balanced sentences, again only 292 utterances can provide a recognition rate (88.75%) even slightly higher than that given by the 408 manually selected utterances (88.34%) at a much higher learning slope. This is because the 292 phonetically distributed syllable utterances cover the top 100 most frequently used tonal syllables out of 1345, which is very helpful in improving the overall performance. Thus the statistical distributions for the acoustic units have made the adaptation process much more efficient. Similarly, the third stage with a total of 436 phonetically distributed utterances requires only slightly more utterances than the 408 manually selected syllables, but provides an accuracy (92.18%) comparable with that obtained by 1345 × 2 tonal syllables (92.28%). When the four stages of phonetically distributed sentences with a total of 992 syllable utterances (significantly less than a single set of the 1345 tonal syllables) are all entered, the recognition rate achieved (93.64%) is only slightly lower than the rate (95.08%) obtained by 1345 × 3 adaptation utterances (which is more than four times of 992 utterances). These results clearly verified the efficiency of the proposed scheme for automatically selecting sentence sets for adaptation purposes as mentioned previously.

The above four-stage adaptation procedure using phonetically distributed sentence sets has actually been used in a very successful Mandarin dictation prototype system developed in early years, Golden Mandarin(II) (Lee *et al.*, 1993*b*; Lee, 1997). In practice, because the incremental adaptation can be implemented as an on-line process in real time, a new speaker does not have to complete all the four stages before using the system. He can decide to begin

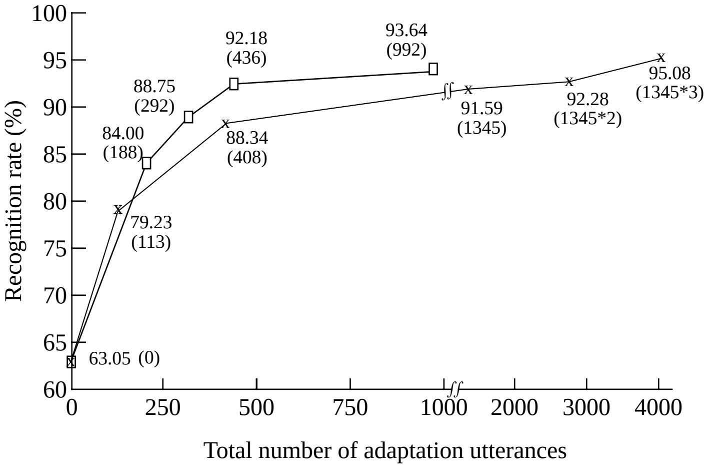

Figure 3. The learning curves for the adaptation procedure for isolated syllables using phonetically distributed sentences (—□—) and manually selected utterances (—×—)

to use the system directly at any time because further adaptation can always be performed on-line during real applications as long as corrections can be made. Similarly such on-line adaptation can also be performed after the four-stage adaptation is completed. This on-line process after the four-stage adaptation can continuously improve the performance slightly, and eventually become saturated.

7. Experimental results for continuous speech recognition

The next set of experiments were performed for continuous speech recognition using the second set of speech database produced in continuous speech mode as presented in Section 5, including those for the three incremental sets of phonetically distributed sentences obtained in Section 3.2. The results are discussed in this section. Here the left-to-right continuous density HMM(CHMM) was used to model the 113 CD INITIAL/41 CI FINAL units (Lee *et al.*, 1993*a*). The conventional one-pass Viterbi beam search algorithm with a fixed pruning threshold was used to decode the optimal syllable sequence in a continuous Mandarin sentence (Lee *et al.*, 1992).

7.1. *Adaptation results for the selected phonetically distributed sentence sets*

The testing data produced by the testing group B1 using the testing texts of six paragraphs covering six different subject domains as described in Section 5 are used here. As shown in Table IV, the average base syllable recognition rate for the testing group B1 using the speaker-independent (SI) models trained from the database produced by the speaker group B2 is 55.57%. Also, significant variations on the recognition rates across the three test speakers can be observed, ranging from 49.59% to 62.83%. Now with the incremental phonetically distributed training sentence sets obtained in Section 3.2 produced by the testing group B1, the SI models were then adapted to the new user stage-by-stage just as in the isolated syllable recognition case. As can be seen in Table IV, the average recognition rates for base syllables can be immediately improved from 55.57% to 69.50% after the first stage using only 50 s of speech data with the simplified on-line adaptation algorithm listed as experiment 1 in the

TABLE IV. The three-stage adaptation results for continuous speech recognition using the three sets of phonetically distributed sentences

	Total length of speech signal from stage 1	Experiment 1				Experiment 2
		Speaker 1	Speaker 2	Speaker 3	AVE	AVE
SI results	—	54.28%	49.59%	62.83%	55.57%	55.57%
Stage 1	50 s	63.90%	69.71%	74.89%	69.50%	70.65%
Stage 2	3.1 min	76.29%	77.78%	82.18%	78.75%	79.35%
Stage 3	6.8 min	78.92%	80.00%	85.11%	81.34%	81.85%

left part of Table IV, and then to 70.65% after the re-estimation of model parameters using Bayesian adaptation listed as experiment 2 in the right-hand column of the table. Because all INITIAL/FINAL acoustic units have been covered in the first stage of the 24 training sentences with a desired statistical distribution, the performance can be improved significantly and efficiently. Then with the additional 76 sentences of the second stage covering all the 408 base syllables and the top 20 inter-syllable context dependency classes, the syllable accuracy can be further improved to 78.75% and 79.35% in experiments 1 and 2 respectively, and finally to 81.34% and 81.85% respectively using a total of 6.8 min of speech data up to the third stage, in which more than 96.34% of tonal syllables and more than 92.51% of inter-syllable context dependency classes in the text corpus defining the desired task are covered. Note that the average error rate reductions are more than 59% (44.43% to 18.15%) after the three-stage adaptation process using only 6.8 min of speech data.

Just as in the isolated syllable case, the recognition rates can be further improved with on-line adaptation to some extent and become saturated eventually when the new user actually uses the system after the three stages of adaptation, if he can correct the recognition errors on-line. Also, this three-stage adaptation procedure using the three phonetically distributed sentence sets has been used in a very successful Mandarin dictation prototype system, Golden Mandarin (III) Windows 95 version (Lee, 1997).

7.2. *Adaptation results with respect to different subject domains of the testing data*

It is well known that the accuracies for continuous speech recognition depend heavily on the subject domains for the testing data. Texts on different subjects domains have different vocabularies, different word frequencies and different sentence patterns. Not only the lexicons and language models used to decode the output words and sentences can be completely different, but the statistical distributions for the acoustic units can be quite different as well for texts with different subject domains. The former (differences in lexicons and language models) is beyond the scope of this paper, but we will try to investigate the latter (differences in acoustic unit distributions) here, which has rarely been discussed in the literature. Although the desired task addressed here is very large vocabulary speech recognition with a general domain, as mentioned in Section 3, the majority of the text corpus defining this task is business news, political news and society news. Articles describing all other subject domains are of much less quantity, although they are all present in the text corpus. As a result, the first three out of the six paragraphs of texts used in generating the testing data mentioned in Section 5 (with subject domains of business, politics and society) are much closer to the text corpus defining the desired task, while the last three paragraphs (with subject domains of philosophy, science and sports) are more or less different.

With the above backgrounds, the respective base syllable recognition accuracies for both the speaker-independent models as well as the speaker-adapted models using the above three stages of adaptation data for each of the six test paragraphs with different subject domains are shown in Figure 4. Also shown in Figure 4 are the similarity measures *R* defined in Equation (1) of Section 2 for the acoustic unit distributions between each of the six paragraphs of testing texts and the whole text corpus defining the desired task. The acoustic unit used in evaluating these similarity measures *R* is the base syllable. One can find that when using the speaker-independent models, comparable recognition rates, all relatively low in any case, were obtained for the six test paragraphs despite their different subject domains. As also shown in Figure 4, the similarity measure values *R* evaluated with respect to the whole given text corpus change significantly from 0.4895 to 0.6772 for the six test paragraphs, which indicates the crucial variations in the base syllable distributions for different task domains. Moreover, it can also be observed that the first three test paragraphs are actually much more similar to the given text corpus than the last three test paragraphs in terms of the base syllable distribution, with the average similarity measure values 0.6602 for the former (first three paragraphs) and 0.5669 for the latter (last three paragraphs). Note that in any case all the similarity measure values *R* are relatively low (i.e. much less than unity) even for the first three paragraphs with the right subject domains, apparently because here each of the six paragraphs are not long enough to construct a reasonably good statistical distribution. The results after the three-stage adaptation, however, show that the difference in the recognition rates across the testing paragraphs on different subject domains is, in fact, relatively small, ranging from 79.88% to 84.42%. The average of the first three test paragraphs (83.45%) is about 3.22% higher than the average for the last three test paragraphs (80.23%). It is believed that this feature is due to the very good design of the three adaptation sentence sets. Note that in the second sentence set for the second stage adaptation, all the 408 base syllables are covered regardless of the differences in base syllable distribution for different subject domains. So all different subject domains have been taken care of to some extent in any case. On the other hand, the most frequently used tonal syllables are well covered in the third sentence set in the third stage adaptation, and the most frequently occurring inter-syllable context dependency classes are well covered in the second and third sentence sets in the last two stages of adaptation, both with reproduced statistical distribution, so many frequently used domain-independent words and phrases including function words are all considered as well. In this way, those tasks well defined by the given text corpus can be recognized very well, while the recognition rates for the texts with rather different subject domains will not be too bad as well, and in those latter cases (rather different subject domains) the performance can still be continuously improved after the three-stage adaptation anyway with on-line adaptation algorithm just as in Section 7.1. In other words, the proposed phonetically distributed sentence sets as given in Section 3.2 are especially designed to be used as very good adaptation data for large vocabulary Mandarin speech recognition, even with the difficulties in the variety of subject domains in the test data.

One may wonder what the similarity measure value *R* is between the three sets of phonetically distributed sentences selected for continuous speech recognition obtained in Section 3.2 and the whole text corpus given. This number is found to be 0.8742, significantly higher than any of the numbers in Figure 4, when the base syllable is used as the acoustic unit. This result is quite natural, because the similarity measure value *R* is the criterion used in selecting these sentences in the sets. However, this number is still far from unity, because quite several different phonetic criteria, including those for INITIAL/FINALs, for tonal syllables, and for inter-syllable context dependency classes, instead of that for base syllables, have been used in selecting those sentence sets. We may try to add the last three test paragraphs on quite different subject domains (philosophy, science and sports) to these three sets of phonetically distributed sentences, which results in a slight increase in this similarity measure value *R* from 0.8742

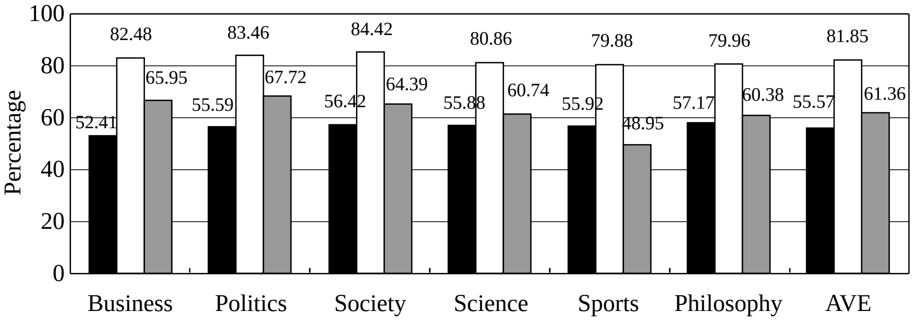

Figure 4. The average recognition rates for speaker-independent and speaker-adapted models as well as the similarity measure values R for the six testing paragraphs with different subject domains. ■, speaker-independent; □, speaker-adapted; ▨, similarity measure value $R(\times 100)$.

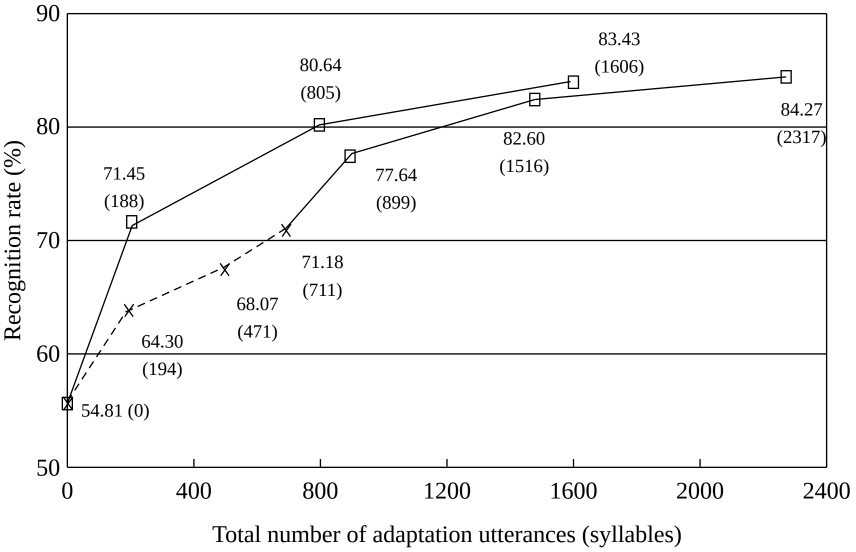

Figure 5. The learning curves using phonetically distributed sentences (—□—) and utterances not matched to the desired task (—×—) for the first three testing paragraphs matched to the desired task.

to 0.8820. In other words, the additional three paragraphs (65 sentences or 711 syllables) on quite different subject domains are not very helpful in improving the similarity of the selected sentence sets with the given text corpus. This is also very natural. As a result, these three paragraphs can provide only very limited improvements in the adaptation processes if the test data are on quite different subject domains.

A series of experiments were further performed to illustrate the above point, in which only the first three paragraphs of test data with subject domains of business, politics and society were tested, while the speech data for the last three paragraphs on quite different subject domains (philosophy, science and sports) were used as the first set of adaptation data, and the data for the three selected sentence sets were used as the second set of adaptation data. The results are plotted in Figure 5, where the lower curve is for the first set of adaptation data of three paragraphs with different subject domains, and the upper curve is for the second set of adaptation data with the selected sentence sets, both started with the same speaker-independent models as used previously trained by the data produced by the speakers in group B2. Note that the data for the upper curve are very similar to those for the three stages in Table IV, but slightly higher. This is because the data in the upper curve in Figure 5 are for only the three paragraphs of test texts with subject domains closer to the given text corpus, but the data in Table IV also include results for the other three paragraphs with different subject domains. It is noted that the learning slope using the first set of adaptation data for the three different subject domains (the lower curve) is always much lower than that using the second set of adaptation data for the three selected sentence sets (the upper curve). For example, although the recognition rates can be improved from 54.81% to 64.30% when the first paragraph with subject domain of philosophy in the first set of data is used (23 sentences or 194 syllables), this number is much worse than the result using the first set of phonetically distributed sentences, which is 71.45% using 188 syllables. After the other two paragraphs on science and sports in the first set of adaptation data were added (with a total of 65 sentences or 711 syllables), the finally achieved recognition rate (71.18%) is still much lower than that obtained by the first two stages of the selected phonetically distributed sentence sets (80.64% using 805 syllables). So the learning slope using some adaptation data not matched to the desired task is always much lower than that using the selected phonetically distributed sentences matched to the desired task. Now in the next set of experiments, in which the second set of adaptation data for the three phonetically distributed sentence sets were further used as the adaptation data applied on the models previously adapted by the three paragraphs of different subject domains with accuracy 71.18% as shown in the lower curve in Figure 5. The recognition accuracy for this case can then be improved to 77.64%, 82.60% and 84.27% respectively. These results are plotted in Figure 5 as the lower curve (full lines) connecting to the dashed line. It can be found that this curve still has a relatively high learning slope, but the final result of 84.27% after all the three sets of selected sentences were used is now only slightly better than that achieved in the upper curve (83.43%) using only the three sets of selected phonetically distributed sentences. In other words, the additional three paragraphs of adaptation data on different subject domains actually provides only very limited overall improvements on recognition performance. All these results again verified the efficiency of the approach proposed in this paper, as well as the point that the adaptation data dynamically selected for a desired task domain always gives a far faster and more efficient speaker adaptation.

8. Conclusion

In this paper, we have investigated the approach of automatic selection of phonetically distributed sentences for fast and efficient speaker adaptation. Incremental sets of different phonetically distributed sentences can be selected automatically based on different acoustic criteria, and they can then be integrated into a multi-stage adaptation procedure. Although the complete experiments were performed for Mandarin syllable recognition, the concepts and techniques here are believed to be equally applicable to many other speaker adaptation tasks in different languages.

References

Cox, S. J. & Bridle, J. (1995). Predictive speaker adaptation in speech recognition. *Computer Speech and Language* **9**, 1–17.

Digalakis, V. & Neumeyer, L. (1995). Speaker adaptation using combined transformation and Bayesian methods. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 680–683.

Furui, S. (1989). Unsupervised speaker adaptation based on heirarchical spectral clustering. *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**, 1923–1930.

Gauvain, J. L. & Lee, C. H. (1994). Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Signal Processing* **2**, 291–298.

Hao, Y. & Fang, D. (1994). Speech recognition using speaker adaptation by system parameter transformation. *IEEE Transactions on Speech and Audio Processing* **2**, 63–68.

Hattori, H. & Sagayama, S. (1992). Vector field smoothing principle for speaker adaptation. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 381–384.

Huang, X. (1992). Speaker normalization for speech recognition. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 465–468.

Huang, X. & Lee, K. F. (1993). On speaker-independent, speaker-dependent, and speaker-adaptative speech recognition. *IEEE Transactions on Speech and Audio Processing* **1**, 150–157.

Huo, Q. & Lee, C. H. (1997). On-line adaptative learning of the continuous density hidden Markov model based on approximation recursive Bayes estimate. *IEEE Transactions on Signal Processing* **5**, 161–172.

van Santen, P. H. Jan, & Buchsbaum, A. L. (1997). Methods for optimal text selection. *Proceedings of Eurospeech'97*, pp. 553–556.

Kohonen, T. (1988). Learning vector quantization. *Abstracts of the First Annual International Neural Network Society Meeting* **1**, 303.

Kosaka, T., Matsunaga, S. & Sagayama, S. (1996). Speaker-independent speech recognition based on Tree-structured speaker clustering. *Computer Speech and Language* **10**, 55–74.

Lee, L. S. (1997). Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine* **14**, 63–101.

Lee, C. H., Giachin, E., Rabiner, L. R., Pieraccini, R. & Rosenberg (1992). Improved acoustic modeling for large vocabulary continuous speech recognition. *Computer Speech and Language* **6**(2), 103–127.

Lee, C. H., Lin, C. H. & Juang, B. H. (1991). A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Transactions on Signal Processing* **39**, 806–814.

Lee, L. S., Tseng, C. Y., Gu, H. Y., Liu, F. H., Chang, C. H., Lin, Y. H., Lee, Y. M., Tu, S. L., Hsieh, S. H., Chen, C. H. (1993*a*). Golden Mandarin(I) — A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary. *IEEE Transactions on Speech and Audio Processing* **1**, 158–179.

Lee, L. S., Tseng, C. Y., Chen, K. J., Hung, I. J., Lee, M. Y., Chien, L. F., Lee, Y. M., Lyu, R. Y., Wang, H. M. (1993*b*). Golden Mandarin(II) — An improved single-chip real-time Mandarin dictation machine for Chinese language with very large vocabulary. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 503–506.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech and Language* **9**, 171–186.

Lyu, R. Y., Hong, I. C., Shen, J. L., Lee, M. Y. & Lee, L. S. (1998). A new approach for isolated Mandarin syllable recognition based upon segmental probability model (SPM). *IEEE Transactions on Speech and Audio Processing* **6**, 293–298.

Matsukoto, H. & Inoue, H. (1992). A piecewise linear spectral mapping for supervised speaker adaptation. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 449–452.

Shen, J. L., Wang, H. M., Lyu, R. Y. & Lee, L. S. (1994). Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models. *Proceedings of International Conference on Spoken Language Processing*, pp. 443–446.

Shikano, K., Lee, K. F. & Reddy, R. (1986). Speaker adaptation through vector quantization. *Proceedings of International Conference on Acoustics, Speech Signal Processing*, pp. 2643–2646.

Stern, R. M. & Lasry, M. J. (1987). Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**, 751–763.

Takahashi, S. & Sagayama, S. (1997). Vector-field-smoothed Bayesian learning for fast and incremental speaker/telephone-channel adaptation. *Computer Speech and Language* **11**, 127–146.

Wang, H. M., Ho, T. S., Yang, R. C., Shen, J. L., Bai, B. R., Su, J. H. & Lee, L. S. (1997). Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. *IEEE Transactions on Acoustics, Speech, Signal Processing* **5**.

Shiow-min Yu & Chi-shi Liu (1990). The construction of phonetically balanced Chinese sentences. *Telecommunication Laboratories Technical Journal*, R.O.C. **28**, 84–91.

Zhao, Y. (1994). An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, **2**, 380–394.

(Received 09 March 1998)