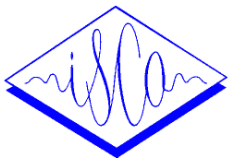

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Incremental Speaker Adaptation Using Phonetically Balanced Training Sentences for Mandarin Syllable Recognition Based on Segmental Probability Models

Jia-lin Shen, Hsin-min Wang, Ren-yuan Lyu and Lin-shan Lee
Dept. of Electrical Engineering, Rm. 520, National Taiwan University
Taipei, Taiwan, R.O.C.

Abstract

This paper presents a new incremental speaker adaptation technique for isolated Mandarin syllables using four sets of phonetically balanced sentences. This algorithm was based on a newly developed Segmental Probability Model(SPM) which was found specially suitable for isolated Mandarin syllable recognition. Each Mandarin syllable is conventionally divided into INITIAL and FINAL parts and based on the INITIAL/FINAL structure of Mandarin syllables, a segment sharing concept is first proposed. A computer algorithm was then developed to select automatically four sets of phonetically balanced sentences with different selection criterion from a large Chinese text corpus. After the four-stage adaptation procedure, the recognition rate for a new speaker can be improved from 63.05% to 92.96%.

1 Introduction

Mandarin Chinese is a monosyllabic-structured tonal language. There exists a total of 1345 different syllables only. When the differences in tone are disregarded, these 1345 different syllables are reduced to 408 base syllables. Because 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, since every Chinese character is simply pronounced as a monosyllable. Various speaker adaptation techniques have been proposed[1][2], including either spectral mapping or model transformation approaches. In our recent results for Mandarin dictation task[3][4], Segmental Probability Model(SPM) was found to be very useful for isolated Mandarin syllable recognition. 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[5]. The average error rate for these 408 base syllables in speaker dependent mode trained by 3 sets of all the 1345 syllables using SPM is 7.11%. However, when a speaker independent model trained by 70 speakers is used, the error rate is increased to 40.40%, more than 5 times higher than that of the speaker dependent result. Therefore efficient speaker adaptation techniques are necessary for new speakers to train the system from the initial speaker independent models.

Conventionally each Mandarin syllable is decomposed into an INITIAL/FINAL format, where INITIAL means the initial consonant of the syllable and FINAL means the vowel(or diphthong) part but including possible medial and nasal ending. There are a total of 22 different INITIAL's and 41 different FINAL's. The 22 different context-independent (CI) INITIAL's can be further expanded to 113 context-dependent(CD) INITIAL's in isolated syllables considering the beginning phoneme of the following FINAL's. A segment sharing concept for SPM is thus proposed in this paper, in which the first few segments(or states) of the SPM's for syllables having the same CD INITIAL's bear similar characteristics so can share the same training data, and so do the remaining segments of the models for syllables having the same FINAL's. In this way the necessary training data for a new speaker can be significantly reduced.

On the other hand, a computer algorithm was also developed to select automatically four sets of phonetically balanced training sentences from a large Chinese text corpus. Each of such sets not only includes almost minimum numbers of sentences covering all necessary phonetic units, each with a different selection criterion, but has the statistical distributions for selected phonetic units very close to those in the large text corpus. In this way the frequently used units can be better trained with higher accuracy, but the new user needs only read smallest number of meaningful sentences to train the system. The first stage of adaptation needs a phonetically balanced sentence set of only 188 syllables(characters) or 24 sentences which covers all the 113 CD INITIAL's and 41 FINAL's. This set of sentences can improve the recognition rate of a new speaker from 63.05% immediately to 82.28%. The following three sets of phonetically balanced sentences, on the other hand, include 104, 144 and 556 characters but covers the top 100, 200 and 500 most frequently used syllables out of 1345, and these syllables in fact cover 48.82%, 69.24% and 93.83% of syllables in daily used Mandarin Chinese respectively. They can further improve the recognition rate of a new user to 87.81%, 90.19% and 92.96% respectively.

This paper is organized into 4 sections. In section 2, the segment sharing concept is discussed and an incremental speaker adaptation process based on this concept is proposed. The computer algorithm to select phonetically balanced sentences and the experimental results for speaker adaptation are discussed in section 3. Section 4 makes the concluding remarks.

2 Segment Sharing Concept

As mentioned above, conventionally each Mandarin syllable is decomposed into an INITIAL/FINAL format. The total numbers for context-independent(CI) INITIAL's and CI FINAL's are 22 and 41 respectively. These 41 CI FINAL's can be further classified into 8 groups according to their beginning phonemes. For example, /a/, /an/, /au/, /ang/, and /ai/ can be classified into same group because they have the same beginning phoneme /a/. The FINAL's in the same group is assumed to have the same coarticulation with their preceding INITIAL's, for example, the /b/ part of /ba/, /ban/, /bau/, /bang/, /bai/... etc. are assumed the same. However, the INITIAL's with the following FINAL's in different groups are regarded as different INITIAL's, for example, the /b/ part of /ba/, /bei/, /bi/, /bu/, /bo/... etc. are assumed different. In this way, the 22 INITIAL's can therefore be further expanded to 113 CD INITIAL's.

The segment sharing concept for SPM is then proposed based on the above CD INITIAL/CI FINAL structure of Mandarin syllables. That is, the first few segments(or states) of the SPM's for syllables having the same CD INITIAL's actually bear similar acoustic characteristics thus can share the same training data, and so do the remaining segments of the models for syllables having the same CI FINAL's. This is because the first few segments primarily include the INITIAL part plus the transition part, while the remaining segments primarily model the FINAL part.

2.1 Initial Experiments

The first experiment performed is to select the number of segments, N. A set of speaker dependent segment-shared SPM's were trained for each of three speakers. For each speaker every syllable model is trained by 3 sets of 1345 syllables, so the total number of training utterances for each speaker is 1345×3 = 4035. The number of segments N tested are 2,3,4 and 5 respectively. The test database contains a fourth collections of the 1345 syllable utterances produced by each of the above 3 speakers. Because the INITIAL part is very short compared with the whole syllable, it is reasonable to assume that the number of segments which model the CD INITIAL part cannot be larger than that of segments which model the FINAL part. For example, the possible partition for N = 4 is (1,3) and (2,2) where the first integer represents the number of segments modeling the CD INITIAL part, and the second integer represents that of the FINAL part. The results for the 3 speakers listed in Table 1 show that N=3 with (1,2) partition provides the highest recognition rate (92.95% averaged over the 3 speakers), which even slightly exceeds the result of the original syllable models(i.e., no segment sharing, simply 408 models) trained by 4035 syllable utterances as shown in Table 2(averaged 92.89% for the same speakers). Table 2 in fact compares the speaker dependent case for N=3, (1,2) partition as in Table 1 and original models without segment sharing with respect to a speaker independent case trained by 70 speakers and tested by the 3 outside speakers. Apparently the speaker independent case is much worse. Note that in the speaker dependent case of Table 2, the segment shared model is only slightly higher than the original model. This is believed due to the availability of enough training data. When only 1 or 2 sets of training utterances are available, the segment shared models perform much better as shown in Fig.2. It can be seen that the segment sharing concept can improve the recognition rate from 73.65% to 86.64% when only one set of 1345 training utterances are available, and from 85.75% to 90.42% when two sets of training utterances are available. Apparently, the recognition rate improvements degrade as the training data increases. This is because when the segment sharing concept is applied, the number of available training utterances for each segment is increased in average by more than 6 times that of the original models. However, when more training utterances are available, the averaging effect in sharing of uniform segments from different syllables also makes the models more coarse or less accurate. This is why the improvements provided by segment sharing eventually saturated. It can also be noted that in the speaker independent results in Table 2, the segment sharing concept can slightly improve the recognition rate from 59.50% to 63.05%. This is because the differences between speakers can be kind of compensated for when more training data can be available.

			Sp1	Sp2	Sp3	AVE
shared model	N=2	(1,1)	84.23	90.05	85.78	90.02
	N=3	(1,2)	93.92	92.39	92.53	92.95
	N=4	(1,3)	93.86	87.40	91.33	90.86
		(2,2)	93.11	94.46	91.25	92.87
	N=5	(1,4)	89.71	85.27	92.02	89.00
		(2,3)	92.92	92.90	91.95	92.59

Table 1: The recognition results for segment shared models with number of segments varying from 2 to 5

2.2 Speaker Adaptation

Based on the concept of segment sharing mentioned above, only very limited training data for a new speaker covering all the 113 CD INITIAL's and 41 FINAL's can be used to adapt all the parameters of SPM's. Therefore 115 syllables are first manually selected as the adaptation data by the following rules :

1. every CD INITIAL/CI FINAL must appear at least once.

2. the total number of raining utterances should be minimum.

For each segment in SPM, because the speech features are modeled by Gaussian mixtures, the adaptation process includes only the reestimation of mean vectors and covariance matrices of the Gaussian distributions. In this study, only diagonal covariance matrices are considered for the Gaussian distributions. Given the speaker independent model, a simple speaker adaptation algorithm based on interpolation and the segment sharing concept is developed as follows:

1. Equally divide all the adaptation syllable utterances into N segments and label these segments with the CD INITIAL's or CI FINAL's which they belong to.

2. For each segment of the SPM's, collecting all adaptation data.

3. Encode the speech features of the collected adaptation data to all the mixtures of that segment. For each mixture j, an interpolation process is relatively easy:

$$\mu_{j,ad} = \frac{N_j}{N_j + n_j}\mu_j + \frac{n_j}{N_j + n_j}\rho$$

$$\sigma^2_{j,ad} = \frac{N_j}{N_j + n_j}(\mu_j^2 + \sigma_j^2) + \frac{n_j}{N_j + n_j}(\rho^2 + \tau^2) - \mu^2_{j,ad}$$

$$N_j = N_j + n_j$$

where μ_j and σ_j^2 are the mean and variance of the j-th mixture component of that segment in the initial model, while ρ and τ^2 are the mean and variance obtained from the adaptation data collected in that mixture. N_j is the weight which is assigned with an initial value in the beginning of the adaptation process, while n_j is the total number of frames of speech features collected in that mixture.

Table 3 is the experimental results when both the mean vector and covariance matrix are adapted with total number of adaptation syllables varying from 115 up to 4035. It can be found that with only 115 adaptation syllable utterances, the error rate can be reduced by 26.52%. When 408 base syllables are included in the adaptation data, the error rate can be further reduced by 59.27% in comparison with the initial speaker independent results. When 1,2 and 3 collections of the 1345 utterances are used for adaptation, the error rates are reduced by 73.61%, 75.75% and 82.71% respectively in comparison with the speaker independent results. In fact, the recognition rates are even better compared with the speaker dependent case. This is because the information of speaker-independent model and speaker-specific data are combined in the adaptation procedure.

		Sp1	Sp2	Sp3	AVE
SD results	original model	94.32	93.47	90.87	92.89
	SS model	93.92	92.39	92.53	92.95
SI results	original model	58.29	57.17	63.34	59.60
	SS model	59.95	59.26	69.95	63.05

Table 2: speaker-dependent(SD) and speaker-independent(SI) results for original and SS(segment shared) models

3 Adaptation Using Incremental Phonetically Balanced Sentence Sets

Conventionally, generating large number of isolated training data for a Mandarin syllable recognition system is a boring and time consuming procedure because each isolated syllable does not bear any meaning. A good idea is to combine all necessary syllables into meaningful sentences. Since these sentences must include all desired units(CD INITIAL's, CI FINAL's, syllables, etc.), they are in fact some kind of phonetically balanced sentences. In this way, the new user can train the system by simply reading these meaningful sentences character by character, which will be much more interesting and not boring at all.

A computer algorithm has been proposed[6] to select from a large text corpus incremental sets of phonetically balanced sentences with different chosen acoustic units. A total of four phonetically balanced sentence sets are thus chosen to form a four stage adaptation procedure as shown in Figure 1. In the first stage, a phonetically balanced sentence set covering all the 113 CD INITIAL's and 41 CI FINAL's. This set consists of only 24 sentences or 188 syllables(characters). In the second stage, 15 additional sentences or 104 additional syllables(characters) are added to form a phonetically balanced sentence set covering the top 100 most frequently used syllables out of 1345. In the third and fourth stage, additional sentences and syllables or characters are used to cover the top 200 and 500 most frequently used syllables. The corpus used here consists of a total of 124,845 sentences(1,374,182 characters) collected from daily newspapers. In this way, the speech data read for these phonetically balanced sentences can be used as very good adaptation data for a new speaker. After the first stage, the 188 characters uttered by a new speaker already includes all the 113 CD INITIAL's and 41 CI FINAL's. Thus the segment sharing concept introduced above can be applied to modify all the model parameters and the recognition rate can be significantly improved very efficiently. After the second, third, and fourth stages, the new speaker will have to utter only 104, 144 and 556 additional syllables but the top 100, 200, and 500 frequently used syllables in fact cover 48.82%, 69.24%, and 93.83% of syllables in daily used Mandarin Chinese respectively. Furthermore, these sentence sets also reproduce(to a very good approximation) the statistical distribution of the CD INITIAL's, CI FINAL's and syllables in the large text corpus such that the more frequently used units can be trained better and recognized more accurately, thus better overall recognition rate can be achieved.

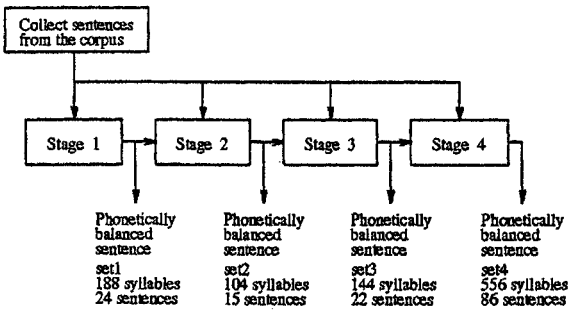

Figure 1: The block diagram of the hierarchy sentence selection algorithm with 4 stages.

3.1 Experimental Results

With the above mentioned incremental speaker adaptation procedure and the four sets of phonetically balanced sentences produced by the new speaker, the system can adapt to a new speaker stage by stage. Table 4 is the experimental results for the four-stage incremental adaptation procedure evaluated for 3 outside speakers respectively. The recognition results are in fact the average accuracy

3. Encode the speech features of the collected adaptation data to all the mixtures of that segment. For each mixture j, an interpolation process is relatively easy:

$$\mu_{j,ad} = \frac{N_j}{N_j + n_j}\mu_j + \frac{n_j}{N_j + n_j}\rho$$

$$\sigma^2_{j,ad} = \frac{N_j}{N_j + n_j}(\mu_j^2 + \sigma_j^2) + \frac{n_j}{N_j + n_j}(\rho^2 + \tau^2) - \mu^2_{j,ad}$$

$$N_j = N_j + n_j$$

where μ_j and σ_j^2 are the mean and variance of the j-th mixture component of that segment in the initial model, while ρ and τ^2 are the mean and variance obtained from the adaptation data collected in that mixture. N_j is the weight which is assigned with an initial value in the beginning of the adaptation process, while n_j is the total number of frames of speech features collected in that mixture.

Table 3 is the experimental results when both the mean vector and covariance matrix are adapted with total number of adaptation syllables varying from 115 up to 4035. It can be found that with only 115 adaptation syllable utterances, the error rate can be reduced by 26.52%. When 408 base syllables are included in the adaptation data, the error rate can be further reduced by 59.27% in comparison with the initial speaker independent results. When 1,2 and 3 collections of the 1345 utterances are used for adaptation, the error rates are reduced by 73.61%, 75.75% and 82.71% respectively in comparison with the speaker independent results. In fact, the recognition rates are even better compared with the speaker dependent case. This is because the information of speaker-independent model and speaker-specific data are combined in the adaptation procedure.

		Sp1	Sp2	Sp3	AVE
SD results	original model	94.32	93.47	90.87	92.89
	SS model	93.92	92.39	92.53	92.95
SI results	original model	58.29	57.17	63.34	59.60
	SS model	59.95	59.26	69.95	63.05

Table 2: speaker-dependent(SD) and speaker-independent(SI) results for original and SS(segment shared) models

3 Adaptation Using Incremental Phonetically Balanced Sentence Sets

Conventionally, generating large number of isolated training data for a Mandarin syllable recognition system is a boring and time consuming procedure because each isolated syllable does not bear any meaning. A good idea is to combine all necessary syllables into meaningful sentences. Since these sentences must include all desired units(CD INITIAL's, CI FINAL's, syllables, etc.), they are in fact some kind of phonetically balanced sentences. In this way, the new user can train the system by simply reading these meaningful sentences character by character, which will be much more interesting and not boring at all.

A computer algorithm has been proposed[6] to select from a large text corpus incremental sets of phonetically balanced sentences with different chosen acoustic units. A total of four phonetically balanced sentence sets are thus chosen to form a four stage adaptation procedure as shown in Figure 1. In the first stage, a phonetically balanced sentence set covering all the 113 CD INITIAL's and 41 CI FINAL's. This set consists of only 24 sentences or 188 syllables(characters). In the second stage, 15 additional sentences or 104 additional syllables(characters) are added to form a phonetically balanced sentence set covering the top 100 most frequently used syllables out of 1345. In the third and fourth stage, additional sentences and syllables or characters are used to cover the top 200 and 500 most frequently used syllables. The corpus used here consists of a total of 124,845 sentences(1,374,182 characters) collected from daily newspapers. In this way, the speech data read for these phonetically balanced sentences can be used as very good adaptation data for a new speaker. After the first stage, the 188 characters uttered by a new speaker already includes all the 113 CD INITIAL's and 41 CI FINAL's. Thus the segment sharing concept introduced above can be applied to modify all the model parameters and the recognition rate can be significantly improved very efficiently. After the second, third, and fourth stages, the new speaker will have to utter only 104, 144 and 556 additional syllables but the top 100, 200, and 500 frequently used syllables in fact cover 48.82%, 69.24%, and 93.83% of syllables in daily used Mandarin Chinese respectively. Furthermore, these sentence sets also reproduce(to a very good approximation) the statistical distribution of the CD INITIAL's, CI FINAL's and syllables in the large text corpus such that the more frequently used units can be trained better and recognized more accurately, thus better overall recognition rate can be achieved.

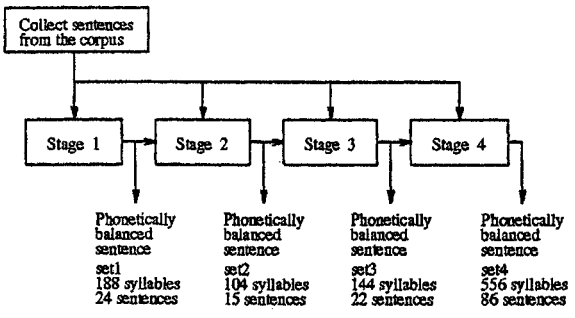

Figure 1: The block diagram of the hierarchy sentence selection algorithm with 4 stages.

3.1 Experimental Results

With the above mentioned incremental speaker adaptation procedure and the four sets of phonetically balanced sentences produced by the new speaker, the system can adapt to a new speaker stage by stage. Table 4 is the experimental results for the four-stage incremental adaptation procedure evaluated for 3 outside speakers respectively. The recognition results are in fact the average accuracy