Isolated Mandarin Syllable Recognition with Limited Training Data Specially Considering the Effect of Tones

Yumin Lee, Lin-Shan Lee, and Chiu-Yu Tseng

***Abstract*—In this correspondence, a set of new approaches is proposed to model the Mandarin syllables for accurate recognition with limited training data while specially considering the effect of tones, including improved initial values and state transition topologies, and making use of the durational cue. The results show that these approaches are very useful practically.**

I. Introduction

The recognition of all the 1300 phonologically allowed isolated Mandarin syllables [1] is a key problem for large vocabulary Mandarin speech recognition. 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, including possible medial and nasal endings. These syllables are highly confusing, because very often many different syllables have exactly the same FINAL but are differentiated only by different INITIAL's. Such syllables form a confusing set, and there exist as many as 38 confusing sets in this vocabulary. This is why the recognition of these syllables is very difficult, and only a speaker-dependent and isolated-syllable task is considered here.

Mandarin Chinese is a tonal language, i.e., every syllable is assigned a tone. There are basically four lexical tones (Tones 1 to 4) and one neutral tone (Tone 5) in Mandarin [2], [3]. It has been shown that the tones can be separately recognized by primarily using pitch contour features. When the differences caused by tones are disregarded, the 1300 different syllables are reduced to 408 *base syllables* (BS's) (which are the syllable structures that carry the pitch contour), and these 408 base syllables can be recognized primarily with vocal tract parameters. Therefore, one can have the BS and the tone of the test utterance separately recognized. An apparent advantage of this approach is that the number of acoustic models required is considerably reduced. In addition, because the tone recognition problem has been successfully solved for isolated syllables in the speaker-dependent case [4], only the BS recognition (but considering the effect of different tones) will be discussed in the rest of this correspondence.

In order to recognize the highly confusing 408 BS's with 38 confusing sets, several special approaches have been proposed [5], [6] and successfully applied to the recognition of the 408 Mandarin syllables with Tone 1 only (i.e., both the training and test utterances are of Tone 1 only). However, the recognition performance of these approaches developed for Tone 1 only will be, as will be shown later in this correspondence, seriously degraded when directly applied to the BS recognition for the 1300 Mandarin syllables with different tones. This is due to the large intraspeaker variations in various feature parameters introduced by the tones. For example, only one BS model [ba] is used for the five tonally variant counterparts [ba-1][1], [ba-2], [ba-3], [ba-4], and [ba-5]. This BS model thus tends to have broad distributions for various parameters due to the large variations in utterance characteristics introduced by the different tones. Several special techniques are thus proposed here in this correspondence. Although these techniques are implemented especially for Mandarin base syllables considering the effect of tones, it is believed that similar concepts are potentially applicable to solve similar problems in speech recognition in other languages.

II. Some Preliminaries

A. Speech Database

The speech database contains six collections of utterances, each produced by two male speakers. Each collection produced by one speaker includes one set of 408 training utterances for each of the four lexical tones plus one set of 22 utterances for the neutral tone. More precisely, each set for one of the four lexical tones contains 408 syllable templates, most of them uttered with that particular lexical tone whenever phonologically allowed but sometimes uttered with another chosen phonologically allowed tone if the syllable cannot be produced in that particular tone.

All the speech data are obtained in an office-like laboratory environment without special sound-proof treatment. They are lowpass filtered, digitized, and end-point detected. The sampling frequency is 9.6 kHz. A Hamming window of length 19.2 ms are applied every 6.72 ms with a preemphasis factor of 0.95. Unless otherwise stated, the training data consist of five collections of utterances in the database, and the remaining collection is used in testing. All the results reported here are the average of the two speakers.

B. The Training Approach for the BS Models

Unless otherwise mentioned, left-to-right continuous density hidden Markov models (HMM'S) were used in this paper with seven states and two transitions per state. The output probability density function (pdf) of a state j is the *partitioned Gaussian autoregressive mixtures* (PGAM) function [7]

$$b_j(o_t) = \frac{1}{M} \max_{m=1,2,\cdots,M} \{b_{jm}(o_t)\} \tag{1}$$

where

o_t	feature vector,
M	total number of mixtures,
$b_{jm}(\cdot)$	mth mixture pdf of state j, which is assumed to be an autoregressive Gaussian distribution.

In this type of mixture density, the feature vector space can be considered to be implicitly partitioned into clusters. Each cluster is defined by an autoregressive Gaussian pdf, and the cluster to which a feature vector o_t belongs is found by a nearest-neighbor criterion [7]. This, in fact, resembles the vector quantization (VQ) operation. Such a VQ analogy will be used in later discussions in this correspondence.

For the recognition of the 408 Mandarin syllables with Tone 1 only (all the training and test utterances are in Tone 1 only), very successful results have been obtained using a special HMM training approach [6], as shown in Fig. 1. One set of training utterances are

Manuscript received November 23, 1993; revised March 27, 1996. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

Y. Lee and L.-S. Lee are with the Department of Electrical Engineering, Rm. 521, National Taiwan University, Taipei, Taiwan, R.O.C.

C.-Y. Tseng is with the Institute of History and Philology, Academia Sinica, Taipei, Taiwan, R.O.C.

Publisher Item Identifier S 1063-6676(97)01734-3.

[1] The transliteration symbols used in this paper are the Mandarin Phonetic Symbols II (MPS II). The numerical values following each syllable denote the tone of the syllable.

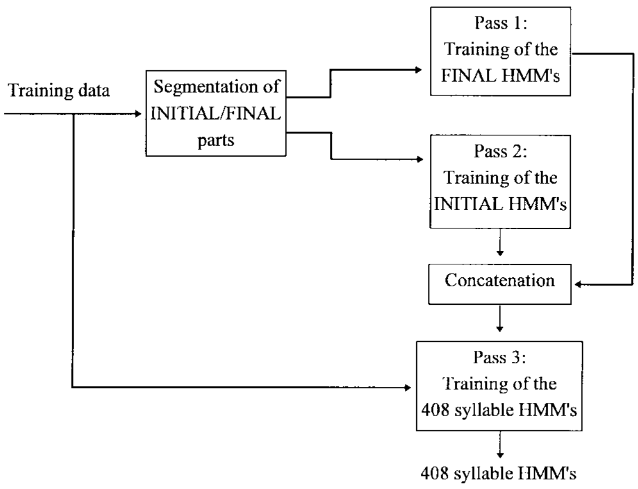
Training data
Segmentation of INITIAL/FINAL parts
Pass 1: Training of the FINAL HMM's
Pass 2: Training of the INITIAL HMM's
Concatenation
Pass 3: Training of the 408 syllable HMM's
408 syllable HMM's

Fig. 1. Block diagram of the original baseline (OBL) training.

first segmented into INITIAL and FINAL parts. These segmented utterances are used to train the INITIAL and FINAL HMM's in the first two passes. They are then concatenated to form 408 syllable HMM's. In the third pass, these 408 syllable HMM's are taken as the initial values, and all of the segmented as well as unsegmented training utterances are used in the Baum–Welch iterations to refine the model parameters, with the parameters for the INITIAL and FINAL parts of the syllable HMM's eventually re-estimated separately. Similar concepts have been extensively used and found very useful in many speech recognition systems for western languages [8]. This approach will be referred to as the original baseline (OBL) approach for further comparison in the rest of this correspondence.

A series of preliminary experiments were performed on the models trained using the above OBL approach but used in the BS recognition for the 1300 Mandarin syllables, i.e., now both the training and test utterances bear all five different tones, but the correct rate is in terms of the correctly chosen BS regardless of the tones. The results of these preliminary experiments for a number of mixtures M ranging from 3 to 6 are listed in the first set of four rows of Table I. One can see that the top BS recognition rates are only on the order of $82{\sim}86\%$, which is significantly lower than those obtained (on the order of 90%) for the recognition of the 408 syllables with Tone 1 only. This is apparently because in the former case, both the training and test utterances bear five different tones, thus having broader (or less discriminating) feature parameter distributions. The top 1 rates are also plotted as the lowest curve in Fig. 2 as a function of M for further comparison later on. Apparently, in this range of M, increased mixtures give higher rates, and $M = 5$ and 6 provide better performance, even with this limited amount of training data.

C. Broadened Tone Coverage

A very intuitive way to improve the performance of the above approach in BS recognition is to improve the initial values used in the Baum–Welch reestimation. In the previous OBL approach, models are initialized by concatenating the INITIAL and FINAL models that are independently trained using the segmented INITIAL and FINAL parts of one set of training utterances bearing Tone 1 only. Therefore, in an improved approach, *four sets* of segmented utterances—*one for each lexical tone*—were used in the first two passes for training the INITIAL and FINAL models to be used as initial values in the final pass of the original approach. Here, only the four lexical tones are considered but not the neutral tone (Tone 5). This is because only 22 syllables, and not 408, can be produced in Tone 5; thus, some special approach will be developed in Section V to improve the BS recognition rates for syllables in Tone 5, and Tone 5 is not considered here. This approach will be referred to as the *broadened tone coverage* (BTC) approach. The top 1 experimental results are plotted as the second lower curve in Fig. 2 for $M = 3$ to 6, and the top 1–5 rates for $M = 5$ and 6 are further listed in the second set of two rows of Table I because the cases $M = 5$ and 6 are very attractive. When compared with the lowest curve in Fig. 2, one can see that an improvement of 0.5% to 2.0% in top-1 rates is obtainable *without changing the training data*. This leads to the new approach of further broadening the tone coverage for the initial values to be presented in the next section.

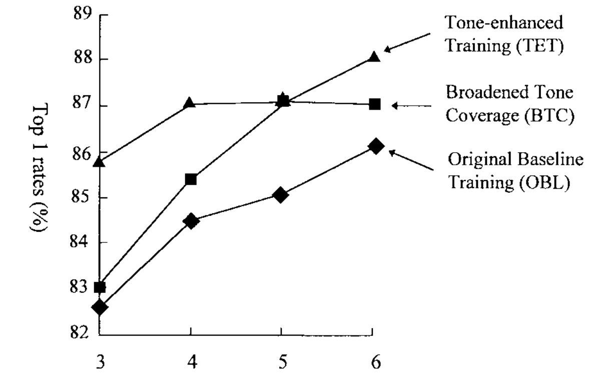

Fig. 2. BS recognition rates for the different training approaches: BSL, BTC, TET.

TABLE I
Top 1 to Top 5 BS Recognition Rates for the Original Baseline Training and Various Improved Techniques with Different Parameters

Approaches			Top *n* rates (%)				
			top 1	top 2	top 3	top 4	top 5
Original Baseline Training (OBL)	M=3		82.71	94.44	96.61	98.31	99.03
	M=4		84.82	95.22	97.40	98.49	98.97
	M=5		85.49	95.28	97.64	98.61	99.15
	M=6		86.58	96.49	98.37	98.91	99.27
Broadened Tone Coverage (BTC)	M=5		87.48	96.01	97.82	98.67	99.15
	M=6		87.42	96.49	98.43	98.91	99.27
Tone-enhance Training (TET) (M_1=M_2=M_4=1)	M=5	M_3=2	87.55	96.25	97.88	98.73	99.21
	M=6	M_3=3	88.51	95.28	96.98	98.13	98.61
Different Topologies (TET, M=5)	A		88.09	95.71	97.58	98.67	99.46
	B		87.79	95.89	97.70	98.85	99.21
	C		87.61	95.83	97.64	98.61	99.33
	D		88.03	95.77	97.58	98.73	99.15
Threshold Decision (TET, Topology A, M=5)	θ=0 (frames)	Neutral	59.09	90.91	95.45	100	100
		Lexial	88.42	95.77	97.61	98.65	99.45
	θ=49 (frames)	Neutral	77.27	100	100	100	100
		Lexial	87.63	94.92	96.75	97.79	98.59

III. The Tone-Enhanced Training Approach

With the above experiences, a new training approach is designed for further broadening the tone coverage of the initial estimates for the BS model parameters, as shown in Fig. 3. In the first two passes, four INITIAL models and four FINAL models are first obtained for each BS, each for one lexical tone. More precisely, for each BS, one INITIAL and one FINAL model, each with M_l mixtures, are obtained using the segmented training data of Tone l, where $l = 1, 2, 3, 4$. These INITIAL and FINAL models are then concatenated to form four syllable models, which are referred to as the component models, each for one lexical tone. Note that at this stage, four component models are obtained for each BS, in which the lth component model having M_l mixtures is trained using a segmented training utterance

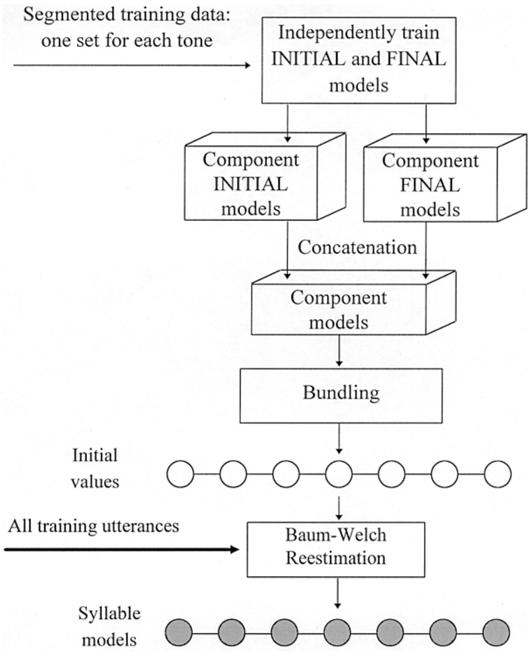

Fig. 3. Tone-enhanced training (TET) approach.

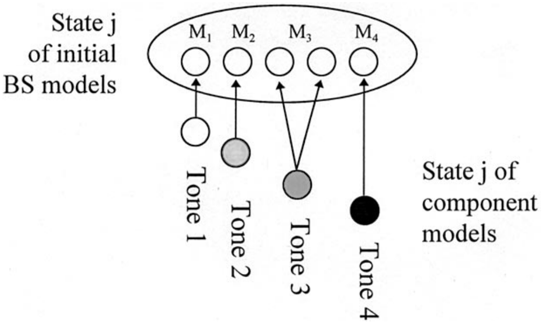

Fig. 4. ”Bundling” operation in the TET approach.

of Tone l. The initial values to be used in the subsequent (or the third-pass) Baum–Welch re-estimation are next obtained by combining the model parameters of the component models. Specifically, for each BS model, the initial state transition matrix **A** is assigned the value of the weighted average of the state transition matrices of the corresponding component models, i.e.,

$$A = \frac{1}{M}\sum_{l=1}^{4} M_l \times A_l \tag{2}$$

where $A_l, l = 1, 2, 3, 4$ denote the state transition matrix of the component model trained for Tone l, and $M \overset{\text{def}}{=} \sum_{l=1}^{4} M_l$ is the number of mixtures of the BS model. The initial state output density to be used in the third-pass re-estimation is then obtained by "bundling" up, or simply copying all of, the state output densities of the four component models, as shown in Fig. 4. In other words, for each state j of the BS model, a total of M initial mixture pdf's are used; they are simply the mixture pdf's copied from state j of the four component models for the four lexical tones. After this "bundling" operation, the obtained BS model is taken as the initial values to go through a third-pass Baum–Welch reestimation process. This approach will be referred to as the *tone-enhanced training (TET) approach* in this correspondence.

Extensive experiments were conducted to investigate the performance of this new approach. The top 1 rates are plotted as the upper curve in Fig. 2 for $M = 3$ to 6, and the top 1–5 rates for $M = 5$ and 6 are also included in the third set of two rows in Table I. The number of mixtures M_l in the component models are, in general, chosen empirically. The fact that M_3 is very often larger than M_1, M_2, and M_4 is based on the observation that the BS recognition rate for the syllables bearing Tone 3 are typically much lower than those of syllables bearing Tones 1, 2, and 4, which is probably due to the special structural variations of the syllables with Tone 3. From these results, one can see that, in general, the top 1 rates of the new approach are on the order of 86%∼88%, which is about 2%∼3% higher than that of the OBL approach and always higher than those obtained by the BTC approach.

An interesting question arises about why this TET approach always performs better than the BTC approach. It was mentioned in Section II that the PGAM state output densities resemble the vector quantization operation in that the feature space is implicitly partitioned into clusters. From this point of view, the training of initial models is in fact analogous to the training of initial VQ codebooks, i.e., the initial partitioning of the feature space. In the BTC approach, the initial model parameters are obtained from the training data in a straightforward way. From the VQ point of view, the initial partitioning of this approach is unsupervised, i.e., it is solely dependent on the distribution of feature vectors, and there is no explicit rule governing the partitioning of the feature space. In the TET approach, however, the initial model parameters are obtained by "bundling" up the parameters of the component models. From the VQ point of view, in this approach, the feature space is initially partitioned in such a way that to a certain extent, the feature vectors of training utterances of the same tone belong to the same cluster. In other words, the initial partitioning of the feature space is no longer unsupervised; there is an explicit rule that assigns M_l clusters to the feature vectors of training utterances of Tone l. This has the effect of equalizing the contributions of the training utterances of different tones.

IV. Special State Transition Topologies

It has been observed that the tones in Mandarin not only influence the distribution of feature vectors but also influence the duration of the syllable utterances [2]. For example, utterances bearing Tone 4 are typically shorter than utterances bearing Tone 1. In other words, the variations in duration can be significant for syllables bearing different tones. Hence, for BS recognition for the 1300 Mandarin syllables, the BS HMM's should be more flexible in duration modeling. The HMM's used in all the previous experiments described above are seven-state left-to-right models with two transitions, as shown in Fig. 5(a), which is referred to as Topology Z. This model may be relatively inflexible in duration modeling for BS recognition. Hence, another possible way of improving the BS recognition performance is to use state transition topologies with more transitions.

Four additional state transition topologies are thus chosen as shown in Fig. 5(b), which are referred to as Topologies A, B, C, and D. Two observations are worth noting in these topologies. First, the number of transitions are considerably increased so that there exist legal paths of model lengths ranging from four to seven states. This has the advantage of better modeling the durational aspects. Second, for each topology, the only legal path from the first three states to the last four states is through the transition from states 3 to 4. This is deliberately designed, considering the fact that the first three states are to model the INITIAL parts and the last four states are to model the FINAL parts.

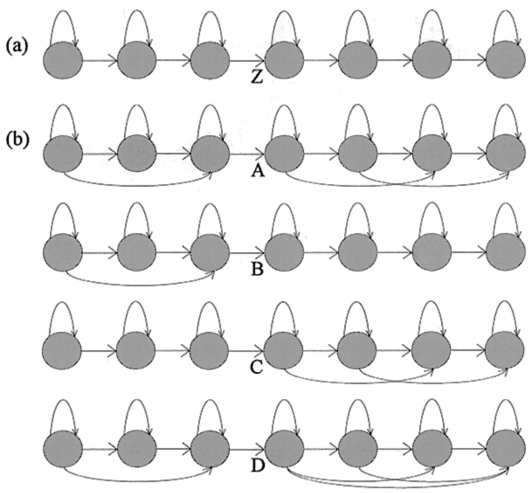

Fig. 5. Different state transition topologies.

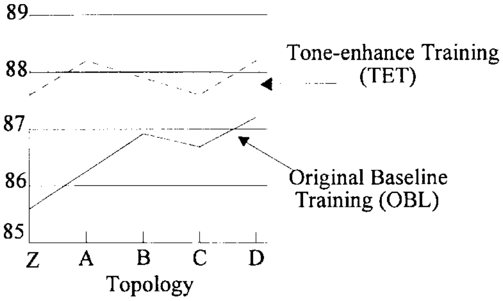

Fig. 6. Top 1 rates using different state transition topologies $(M = 5)$.

A series of experiments were conducted to compare the performance of the different topologies. The number of mixtures was fixed at 5, and both the OBL and the TET approaches with $M_1 = M_2 = M_4 = 1, M_3 = 2$ were used. The top 1 to 5 rates for the TET approach with Topologies A, B, C, D are also listed in the fourth set of four rows in Table I and the top 1 rates for both OBL and TET approaches plotted in Fig. 6. Several interesting observations are worth noting. First, it can be seen that typically, Topologies $A, B, C,$ and D yield higher top 1 rates than Topology Z. The highest top 1 rate of 88.09% now achieved using the TET approach with Topology A is rather close to the 90% performance previously obtained for syllable recognition with Tone 1 only. Second, for each topology, the performance of the TET approach is always better than the OBL approach. Last, Topology A is the most attractive for the TET approach, whereas Topology D is the best choice for the OBL training. Note that Topology D is the most flexible of all the four new topologies in Fig. 5 in terms of the number of allowable state transitions, and Topology A is the second. One possible interpretation for this phenomenon is that for the TET approach, the initial models are very carefully trained; hence, to a certain extent, the effects of the tones on syllable duration have been taken care of. Therefore, such flexibility in state transitions as Topology D is, in fact, unnecessary. For the OBL approach, on the contrary, the initial models are not as elaborate as that for the TET approach, and therefore, the large number of state transitions in Topology D helps to compensate for the variations in syllable duration introduced by the tones.

V. Durational Cue for Syllables with Neutral Tone

The BS recognition for syllables bearing the neutral tone are the most difficult. In fact, the top 1 recognition rates for such cases are usually only on the order of 50–60%, which are significantly lower than the rates for syllables with the four lexical tones, regardless of the training approach and state transition topology used. This is primarily due to the much shorter duration and lower energy in syllables with the neutral tone. Although the number of syllables that can be pronounced with the neutral tone is only 22, practically, the BS recognition for syllables with the neutral tone is really very important. This is because many function words with important syntactic and semantic roles in the Chinese language such as "的", "了", "著", etc., are always pronounced with the neutral tone. These words occur with very high frequency in everyday Chinese, e.g., the occurrence frequency of the function word "的" alone is found to be as high as 7%.

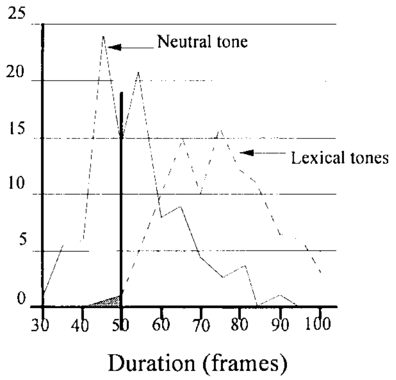

Fig. 7. Distribution of duration in the training utterances for the two types of tones.

The distribution of the durations of the training utterances in the speech database is shown in Fig. 7. One can see that the duration of most syllables with the neutral tone is between 40 to 60 frames, whereas that of most syllables with the four lexical tones is between 60 to 90 frames. One way to make use of this durational cue is by use of a threshold decision rule. Specifically, if the duration of the test syllable is shorter than a threshold θ, then that syllable is assumed to bear the neutral tone and is thus evaluated during the BS recognition process only against the 22 BS models that can be produced with the neutral tone; otherwise, nothing is assumed, and it is evaluated against all the 408 BS models. Since only 22, which is many fewer than 408, syllables can be pronounced with the neutral tone, this procedure can tremendously narrow down the search space and thus significantly improve the performance in BS recognition for syllables bearing the neutral tone with durations shorter than θ. However, for a given θ, syllables shorter than θ but bearing one of the four lexical tones will almost certainly be erroneously recognized because they are only searched through the 22 BS models that can be produced in the neutral tone. Therefore, the BS recognition performance for the syllables with the four lexical tones may be degraded due to such errors, and whether the proposed approach really improves the overall BS recognition performance is contingent on the choice of θ.

Experiments are conducted to test the proposed threshold decision rule with several values of θ. The results are summarized in Fig. 8. Note that the case of $\theta = 0$ is equivalent to the original BS recognition approach, i.e., the recognition approach that does not make any use of the durational cue. The TET approach and Topology A were adopted in these experiments. From the results in Fig. 8, one can see that the top 1 BS recognition rates for the syllables with neutral tone can be considerably improved when the threshold θ increased from 0, yet that for the syllables with the four lexical tones are, in general, only sightly decreased. For example, for the case of $\theta = 49$ frames, the top 1 BS recognition rate for the syllables

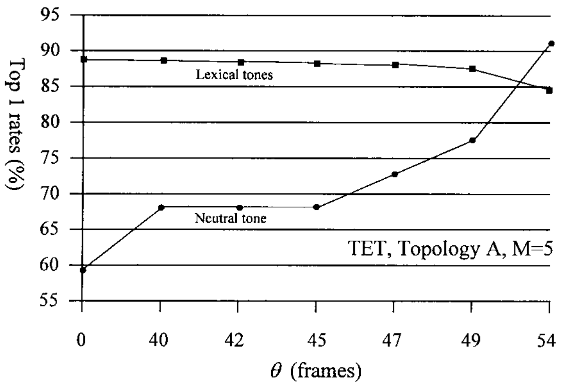

Fig. 8. BS recognition results using the durational cue for different values of θ.

with the neutral tone is improved by 18.18% as compared with the case of $\theta = 0$, yet that for the syllables with the four lexical tones is degraded only by 0.79%. The top 1 to 5 rates for the two types of tones with $\theta = 0$ and 49 frames are also listed in the last set of 4 rows of Table I for comparison.

Although the above results seem to indicate that the proposed threshold approach can significantly improve the correct rate for syllables with the neutral tone but only very slightly degrade the correct rate for syllables with the lexical tones, the actual improvements achievable in the overall BS recognition rate is still difficult to tell, especially when the frequencies of occurrence for the base syllables and tones are considered. The reason is that not all base syllables have the same frequency of occurrence, and the actual percentage of neutral-tone syllables in everyday Chinese is apparently much higher than $22/1300$. This will be investigated in the next section.

VI. Overall Performance Tested on Everyday Chinese

The test database used in all the above experiments include all phonologically allowed Mandarin syllables, and all syllables are equally counted in the performance evaluation. In practical situations, however, some of these syllables are rarely used, whereas others are very frequently used in everyday Chinese, and the ratio of neutral-tone syllables is significantly higher than $22/1300$. Further experiments to test how the concepts proposed in this correspondence actually work in practical situations will be presented here. In these experiments, a large Chinese text corpus including newspapers, articles in magazines, tales, and so on are used to evaluate the frequencies of occurrence of each BS, each tone, and each of the 1300 tonal syllables. The total number of characters (syllables) included in the corpus is around 4 million. Two sets of experimental data are collected: one with the original technique, i.e., the OBL approach and Topology Z and one with the TET approach and the Topology A. Different values of the threshold θ are used. For each case, the BS correct rates for each BS with the two types of tones are weighted and averaged by the corresponding frequencies of occurrence, respectively, and so on; therefore, the results will reflect the performance of the techniques in a real situation.

The results of these experiments are summarized in Fig. 9. Note that the case of $\theta = 0$ is equivalent to the case of making no use of the durational cue. Two observations are noteworthy. First, making use of the durational cue as proposed in Section V can indeed improve the overall top 1 recognition rates, and the optimal value for θ is 42 frames for both cases (i.e., OBL training with Topology Z and TET approach with Topology A). Second, the TET approach with Topology A always yields much higher top 1 rates compared with the OBL approach with Topology Z, regardless of the value of θ. Typical top 1 to top 5 rates for the original case (OBL training, Topology Z, and $\theta = 0$) and the case that all techniques proposed in this paper are integrated (TET approach, Topology A, and $\theta = 42$ frames) are also listed in the top two rows of Table II for comparison. From this table, one can see that the approaches proposed in this correspondence can, in fact, achieve an improvement of 2.61% in the top 1 recognition rate in practical situations, which corresponds to about 20% of error rate reduction.

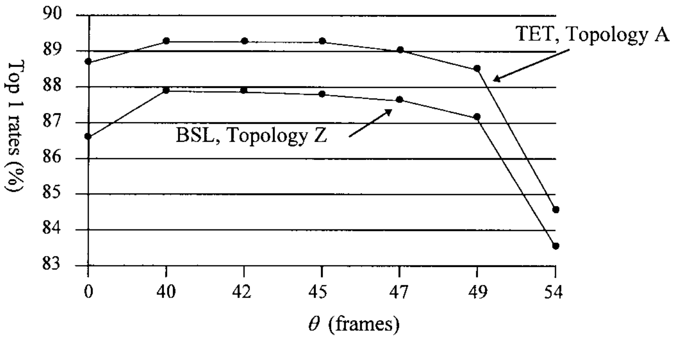

Fig. 9. Test results on everyday Chinese for different values of θ.

TABLE II
Top 1 to Top 5 BS Recognition Rates Tested on Everyday Chinese

Approaches	Top n rates (%)				
	top 1	top 2	top 3	top 4	top 5
OBL, Tolology Z, θ=0 frames	86.63	97.17	98.91	99.24	99.57
TET, Topology A, θ=42 frames	89.24	97.72	98.80	99.13	99.46
1300 models, OBL, Topology Z, θ=0 frames	83.25	94.74	97.22	98.31	99.03

One may wonder that in this correspondence, we started out by saying that the BS's and tones can be separately recognized; we then indicated the problems due to the use of the same model for BS's with different tones and presented a series of techniques to solve these problems. How about if we started by recognizing all the 1300 tonal syllables directly by training 1300 tonal syllable models instead of trying to recognize the BS's and tones separately? Such an experiment was finally performed is and presented here for comparison. A total of 1300 tonal syllable models were trained using the same OBL approach as before including 22 models for the syllables bearing the neutral tone. The only difference is that here the training and testing utterances for each model are of the same tone but can be any one of the five possible tones. In this case, the TET approach, special topology, or the threshold decision approach are all not needed. Exactly the same model structures, i.e., seven states and $M = 5$, left-to-right with two transitions per state, were adopted, and exactly the same training and testing utterances were used. The resulted BS recognition rates—top 1 to top 5—which are also weighted and averaged by the respective frequencies of occurrence just as above, are listed in the last row of Table II. It can be seen that the results are, in general, much lower than the results in the first row of the table, i.e., separately recognizing the BS's and tones but without any improved approaches presented in this paper. Even if some techniques such as improved topologies or the threshold decision may also be applied, less significant improvements are definitely expected because here, different HMM's have been used for syllables with different tones. An apparent reason for the relatively lower performance is due to the limited training data because here, every tonal syllable model is trained by only five (or sometimes six) utterances, which is very small. Although increased training data may be helpful, in the speaker dependent Mandarin syllable recognition problem, a limited database will probably still be a normal situation for some period of time in the future.

VII. Conclusion

A new approach is proposed in this correspondence to obtain more elaborate initial models covering characteristics of different tones. Improved state transition topologies are also found to achieve better performance compared with the simple left-to-right model with two transitions. A threshold decision approach is further developed to improve the performance of BS recognition for the syllables with the neutral tone. The test results on everyday Chinese show that a total error rate reduction on the order of 20% in the top 1 rate can be obtained when all the concepts are properly integrated. Although the techniques here are proposed specially for recognition of Mandarin base syllables considering the effect of tones, it is certainly believed that similar concepts are potentially applicable to solve similar problems in speech recognition in other languages.

References

[1] R. He, Ed., *Guoyurbao Tzdian (Mandarin Chinese Daily Dictionary)*. Taipei, R.O.C.: Guoyurbao, 1976.

[2] L.-S. Lee, C.-Y. Tseng, and M. Ouh-Young, “The synthesis rules in a Chinese text-to speech system,” *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 1309–1320, Sept. 1989.

[3] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: University of California Berkeley Press, 1968.

[4] W. J. Yang, J. C. Lee, Y. C. Chang, and H. C. Wang, “Hidden Markov model for mandarin lexical tone recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 988–992, July 1988.

[5] F.-H. Liu, Y. Lee, and L. S. Lee, “A Direct-concatenation approach to train hidden markov models to recognize the highly confusing mandarin syllables with very limited training data,” *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 113–119, Jan. 1993.

[6] L.-S. Lee *et al.*, “Golden mandarin (I)—A real-time mandarin speech dictation machine for chinese language with very large vocabulary,” *IEEE Trans. Speech Audio Processing*, vol. 1, no. 2, pp. 158–179, Apr. 1993.

[7] B.-H. Juang and L. R. Rabiner, “Mixture autoregressive hidden Markov models for speech signals,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, no. 6, pp. 1404–1413, Dec. 1985.

[8] X. Huang *et al.*, “The SPHNIX-II speech recognition system: An overview,” *Comput. Speech Language*, pp. 137–148, Feb. 1993.

Linear Prediction of the One-Sided Autocorrelation Sequence for Noisy Speech Recognition

Javier Hernando and Climent Nadeu

Abstract— **The aim of this correspondence is to present a robust representation of speech based on AR modeling of the causal part of the autocorrelation sequence. In noisy speech recognition, this new representation achieves better results than several other related techniques.**

I. Introduction

Linear predictive coding (LPC) [1] is a spectral estimation technique widely used in speech processing and, particularly, in speech recognition. However, the conventional LPC technique, which is equivalent to AR modeling of the signal $x(n)$, is known to be very sensitive to the presence of background noise. This fact leads to poor recognition rates when this technique is used in speech recognition under noisy conditions, even if only a moderate level of contamination is present in the speech signal. Similar results are obtained with the well-known mel-cepstrum technique [2]. This explains why some of the main attempts to combat the noise problem consist of finding novel acoustic representations that are more resistant to noise corruption than traditional parameterization techniques.

Linear prediction of the autocorrelation sequence has been the common approach to several robust spectral estimation methods for noisy signals presented in the past. For speech recognition, Mansour and Juang [3] proposed the short-time modified coherence (SMC) as a robust representation of speech based on that approach. On the other hand, Cadzow [4] introduced the use of an overdetermined set of Yule–Walker equations for robust modeling of time series. Although Cadzow applies linear prediction to the signal, his method can also be interpreted as performing linear prediction in the autocorrelation domain. Both methods rely, either explicitly or implicitly, on the fact that the autocorrelation sequence is less affected by broadband noise than the signal itself, especially at high lag indices.

In this work, we consider the one-sided or causal part of the autocorrelation sequence and its mathematical properties. As this sequence shares its poles with the signal $x(n)$, it provides a good starting point for LPC modeling. In this way, the new one-sided autocorrelation LPC (OSALPC) method appears as a straightforward result of the approach [5]. In addition, it is closely related to the SMC representation and Cadzow’s method. All of them can be interpreted as AR modeling of either a spectral function named “envelope” or its square. This interpretation, which is based on the properties of the one-sided autocorrelation, provides more insight into the various methods. In this correspondence, their performance in noisy speech recognition is compared. The optimum model order and cepstral liftering have also been investigated in noisy conditions. The simulation results show that OSALPC outperforms the other techniques in severe noisy conditions and obtains similar scores for moderate or high SNR.

Manuscript received February 14, 1995; revised November 9, 1995. This work was supported by Grant nos. TIC-92-0800-C05/04 and TIC-92-1026-C02/02. The associate editor coordinating the review of this paper and approving it for publication was Dr. Kuldip K. Paliwal.

The authors are with the Department of Signal Theory and Communications, Polytechnical University of Catalonia, Barcelona, Spain.

Publisher Item Identifier S 1063-6676(97)00766-9.

1063–6676/97$10.00 © 1997 IEEE