Tone Recognition of Continuous Mandarin Speech Based on Neural Networks

Sin-Horng Chen, *Senior Member, IEEE*, and Yih-Ru Wang

***Abstract*—Several neural network-based tone recognition schemes for continuous Mandarin speech are discussed. A basic MLP tone recognizer using recognition features extracted from the processing syllable is first introduced. Then, some additional features extracted from neighboring syllables are added to compensate for the coarticulation effect. It is then further improved to compensate for the effect of sandhi rules of tone pronunciation by including tone information of neighboring syllables. The recognition criterion is now changed to find the best tone sequence that minimizes the total risk that simultaneously considers tone recognition of all syllables in the input utterance. Last, two approaches using HCNN and HSMLP, respectively, to model the intonation pattern as a hidden Markov chain for assisting tone recognition are proposed. The effectiveness of these schemes was confirmed by simulations on a speaker-independent tone recognition task. A recognition rate of 86.72% was achieved.**

I. Introduction

Tone recognition is an important task in Mandarin speech recognition because Mandarin Chinese is a tonal language. Each character is pronounced as a monosyllable with which a tone associates. This means the tonality of a monosyllable is also lexically meaningful. Basically, there are only five lexical tones which are commonly labeled in sequence from Tone 1 to Tone 5. The tonality of a monosyllable is mainly characterized by the shape of its fundamental frequency (F0) contour. Although a previous study [1] concluded that the F0 contour of each of the first four tones can be simply represented by a single standard pattern, as shown in Fig. 1, tone recognition in continuous speech is not a trivial problem. This is because tone patterns of syllables in continuous Mandarin speech are subject to various modifications. First, the pronunciation of a Tone 5 is usually highly context dependent so that its F0 contour pattern is relatively arbitrary. This makes Tone 5 difficult to recognize. Second, the tone pattern of a syllable may be seriously affected by the tones of neighboring syllables. This effect is commonly known as sandhi rules [2]. Third, coarticulation effect can make the F0 contour shape of a syllable be affected by the F0 contour patterns of neighboring syllables. Fourth, F0 contour patterns of stressed syllables are generally quite different from unstressed ones. Fifth, the F0 contours of all syllables in a sentential utterance are seriously adjusted to meet the intonation pattern of the sentence. Besides, some other factors, such as the semantics and the emotional status of speaker, can also affect the pronunciations of tones patterns of syllables. Therefore, tone recognition for continuous Mandarin speech is a complicated task. To our knowledge, there have been few studies on this problem [3]–[5]. In this correspondence, we study the problem using neural network-based pattern recognition technique. Several multilayer perceptron (MLP) based tone recognition schemes are proposed to compensate for the effects of some tone modification factors discussed above. They include the coarticulation effect and the sandhi rules from neighboring syllables as well as the effect caused by the intonation pattern of a sentential utterance.

Manuscript received August 30, 1992; revised September 15, 1994. This work was supported by National Science Council under contract no. NSC83-0404-E009-091 and Telecommunications Laboratories, Ministry of Transportation and Communications, Taiwan, R.O.C. The associate editor coordinating the review of this paper and approving it for publication was Dr. Xuedong Huang.

The authors are with Department of Communication Engineering and Center for Telecommunications Research, National Chiao Tung University, Taiwan, R.O.C.

IEEE Log Number 9408401.

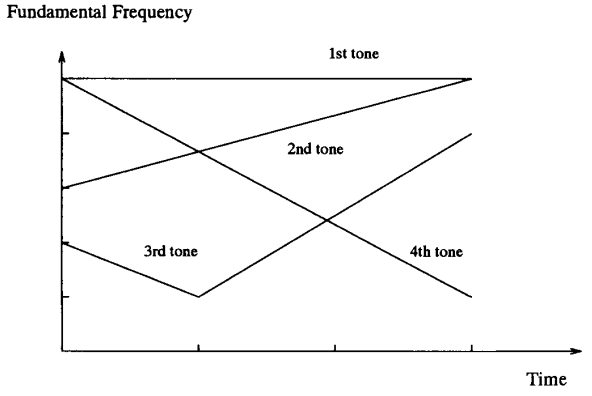

Fig. 1. Standard F0 contour patterns of the first four tones.

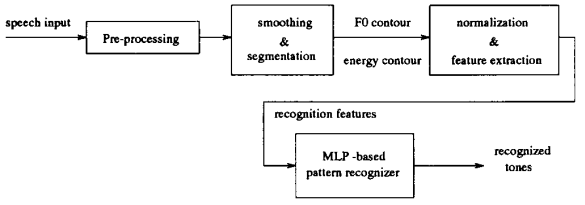

Fig. 2. Block diagram of the basic tone recognition scheme.

II. The Basic Tone Recognition Scheme

Fig. 2 shows the block diagram of the basic tone recognition scheme. It is composed of four parts: preprocessing, smoothing and segmentation, normalization and feature extraction, and MLP-based pattern recognition. The speech signal is the first signal preprocessed to extract some parameters. Preprocessing consists of LP-filtering at 4 kHz, sampling at 10 kHz, A/D conversion into 12-b data, and detection of pitch and log-energy for every 40-ms frame at a rate of 50 Hz. Here, the SIFT algorithm [6] is employed for pitch detection. Because pitch does not exist in unvoiced and silence parts, the pitch contour of each sentential utterance is now divided into segments. A smoothing algorithm operating on a segment-by-segment basis is applied to correct some pitch detection errors. Because a segment may still be composed of pitch contours of several connected syllables, more finely segmenting it into syllable segments is necessary. Due to the fact that finer segmentation of a continuous utterance into syllable segments is more suitably done in the subsystem of recognizing 408 base syllables, we simply do the finer segmentation manually in this study. Finally, all pitch contours of monosyllables are converted into F0 contours.

The F0 contour of each sentential utterance is then normalized by its own means to reduce interspeaker variability. Features for tone recognition are then extracted from the normalized F0 contours and the log-energy contours of syllables. In the basic scheme, only features extracted from the processing syllable were used. They include the duration of the F0 contour of the syllable, the means of three uniformly divided log-energy subcontours, and the

intercepts and slopes of three uniformly divided F0 subcontours [7]. These features are then fed into an MLP pattern recognizer for tone recognition. The MLP is a two-layer network with a single hidden layer. It consists of five output nodes corresponding to five tones. Each neuron output is the sigmoid function of the weighted summation of inputs. The MLP recognizer is trained by the backpropagation (BP) rule [8], which minimizes the mean squared error between the feedforward outputs and the desired targets.

III. The Schemes that Compensate the Coarticulation Effect

As mentioned previously, coarticulation from neighboring syllables will make tone pattern of a syllable change in shape or in level to interfere with tone discrimination. Two types of coarticulation are considered. The first one is due to the continuation of the articulation process. Significance of this type of coarticulation usually depends on the relationship between the processing syllable and its neighbors. Basically, syllables within a polysyllabic phrase are tightly coupled together so that the coarticulation is severe. The F0 contours of these syllables will be distorted in shape so that they can be smoothly transitmitted from one syllable to another. Another type of coarticulation is caused by the sandhi rules [2]. In this case, tone pattern of a syllable is simply affected by tones of neighboring syllables. Sometimes, the tone pattern of a syllable can be totally changed to another tone pattern. For example, a well-known sandhi rule tells that a Tone 3 will be changed to Tone 2 as it follows another Tone 3. Sometimes, the classification of these two types of coarticulation are not clear.

Two tone recognition schemes that compensate for effects of these two types of coarticulation are studied. The first one is a scheme that is directly extended from the basic scheme discussed previously by adding contextual acoustic features extracted from the two nearest neighboring syllables to cope with the first type of coarticulation. It is then further improved in the second scheme by additionally adding tones of the two nearest neighboring syllables as input features to compensate for the effect of sandhi rules. They are discussed in more detail as follows.

In the first scheme, some contextual features extracted from neighboring syllables are additionally fed into the MLP recognizer to help the tone recognition. They include

1) the three features (i.e., log-energy, F0 mean, and slope) extracted from the last subsegment of the preceding syllable
2) the three features extracted from the first subsegment of the following syllable
3) log energies and durations of both unvoiced/silence segments before and after the processing syllable
4) two binary indicators.

Among them, the six features in 1) and 2) are the primary features for coping with the first type of coarticulation. The two features in 3) are used to implicitly represent the tightness of relationships between the processing syllable and the two nearest neighbors. The last two features are, respectively, used to indicate whether the processing syllable is the first or last syllable of the input sentential utterance.

In the second scheme, tones of the two nearest neighboring syllables are also used as the input features to assist the tone recognition. Due to the fact that tones of neighboring syllables are either not known in advance or can only be estimated from previous recognition, tone recognitions for all syllables in an input utterance cannot be done independently. A different recognition procedure based on the decision rule of minimal total risk is therefore employed. First, rather than directly taking the MLP as a tone recognizer, it should be regarded as a mechanism of computing the risk of tone trigram comprising tones of the processing syllable and its two nearest neighbors. Second, define an objective function for the whole input sentential utterance by accumulating risks of all tone trigrams in the utterance. Specifically, given the feature vector sequence $(\underline{X}(j))_{j=1..N}$ of the input sentential utterance, the objective function for the candidate tone sequence $(T(j))_{j=1..N}$ is defined as

$$R((T(j))_{j=1..N}|(\underline{X}(j))_{j=1..N}) = \sum_{j=1}^{N}\sum_{i=1}^{5}\{(O_i(\underline{X}(j),T(j-1),T(j+1)) - t_i(j))^2\}.$$

Here, the output of neural network $O_i(\underline{X}(j),T(j-1),T(j+1))$ is a function of the tone-trigram $(T(j-1),T(j) = \text{tone } i, T(j+1))$, and $t_i(j)$ is the desired output of the ith output neuron for the jth syllable. An optimization procedure is applied to find the best tone sequence $(\hat{T}(j))_{j=1..N}$ that minimizes the objective function. In practical implementation, this can be efficiently accomplished by dynamic programming. We note that a simple linguistic constraint is additionally added in the search of the best tone sequence to inhibit Tone 5 as the tone of the first syllable because it can never happen in natural Mandarin speech.

IV. The Schemes that Cope with the Effect of the Intonation Pattern

Other than the coarticulation effect, the intonation pattern of sentence pronunciation is also an important factor to be considered in tone recognition of continuous Mandarin speech. For instance, the intonation pattern for a declarative sentence generally makes the F0 contour of the utterance decline gradually. Variations on F0 contour caused by tonality as well as other factors are superimposed on the intonation pattern and, hence, make the tone recognition problem of continuous Mandarin speech more difficult. In this section, the tone recognition schemes that cope with the difficulty resulting from the intonation pattern of sentence pronunciation is studied. Two approaches using a hidden control neural net (HCNN) and a hidden state multilayer perceptron (HSMLP) are proposed. The basic idea is to model the global intonation pattern of a sentential utterance as a hidden Markov chain and use a separate recognizer in each state for tone discrimination.

First, the approach using HCNN [9] is discussed. The structure of an HCNN is like an MLP except that it additionally consists of some neurons called hidden control units in the input layer. Rather than connecting to input features, these hidden control units are untouchable and used to generate state-specified information for controlling the neural network to make it vary with time to time align with the input feature sequence. Therefore, an HCNN acts just like a finite-state MLP with outputs depending on both input features and state-specified information generated by hidden control units. In our application, the HCNN is regarded as a sequence of tone recognizers that are time aligned with the hidden Markov chain modeling the intonation pattern. Fig. 3(a) displays the schematic diagram of the HCNN used in our study. It is a left-to-right, three-layer network with single state transition. Hidden control signals assigned to each state are represented by "grandmother-cell" representation, i.e., all bits are equal to '0' except the one associated with the current state being set to '1.'

The training procedure of the HCNN comprises two steps: segmentation and model updating. The first step is to optimally segment the input sentential utterance into several states with the tone sequence known. The second step is to update the HCNN model-based neural network by using the state sequence found from the segmentation result. In recognition phase, the task is to find the best pair of tone and state sequences for the input test utterance using an optimal searching process. An objective function is defined to evaluate the cost for each

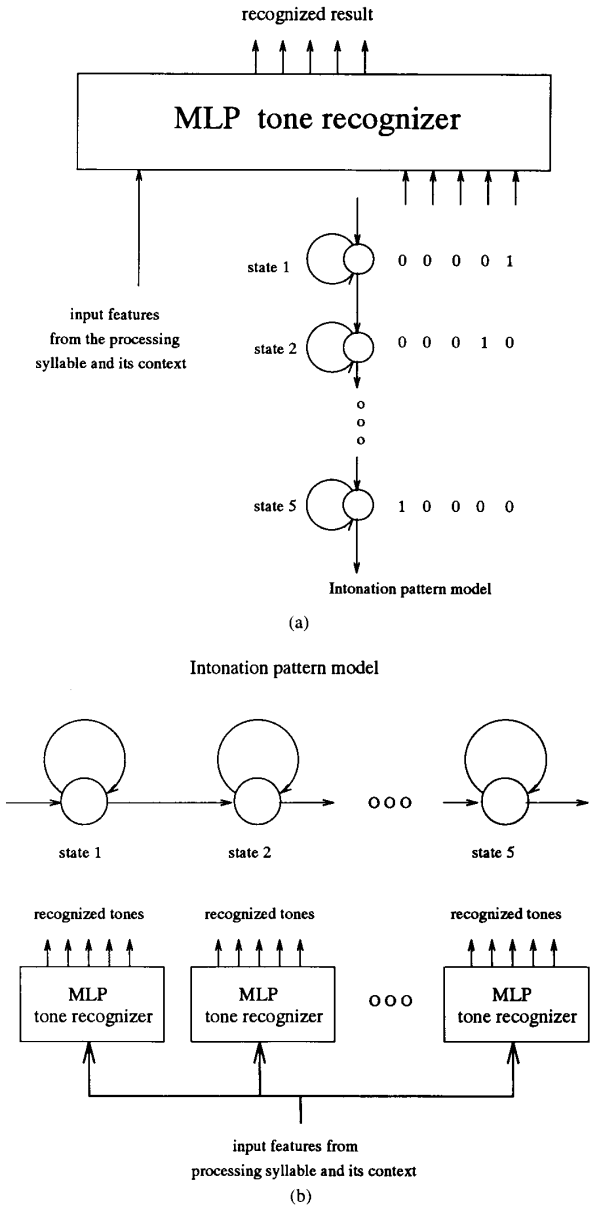

Fig. 3. Schematic diagrams of the schemes using (a) HCNN and (b) HSMLP.

pair of admissible sequence. The criterion is then to find the best pair of tone and state sequences that minimizes the objective function. The following criterion is used:

$$\min_{\mathbf{S},\mathbf{T}} R(\mathbf{S},\mathbf{T}|\mathbf{X})$$

where

$$R(\mathbf{S},\mathbf{T}|\mathbf{X}) = \sum_{n=1}^{N}\sum_{i=1}^{5}\left\{(O_i(\underline{X}(n),T(n-1),T(n+1),s(n)) - t_i(n))^2\right\}$$

is the objective function with $\mathbf{S} = (s(n))_{n=1,N}$, $\mathbf{T} = (T(n))_{n=1,N}$, and $\mathbf{X} = (\underline{X}(n))_{n=1,N}$ denoting the state sequence, the tone sequence, and the input feature vector sequence, respectively. In practical realization, the searching process can be efficiently accomplished by a dynamic programming (DP) procedure.

TABLE I
DISTRIBUTIONS OF FIVE TONES IN THE TRAINING AND TEST SETS

	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5	Total
training set	2778	2766	2238	4168	779	12727
test set	441	485	362	648	97	2033

Due to the fact that feature vectors of two nearest neighboring syllables are also taken as input features, the first and the last syllables of the input utterance are specially treated. Two separate states are used for them. All other states are used for intermediate syllables.

Another approach using HSMLP is now discussed. An HSMLP is composed of a sequence of MLP, which is shown in Fig. 3(b). Each MLP is taken as a tone recognizer for a state of intonation pattern. The structure of the HSMLP is similar to the linked predictive neural net (LPNN) [10] except that all MLP's are taken as tone classifiers instead of predictors. The basic idea of using HCNN and HSMLP to model the hidden Markov chain representing an intonation pattern is the same. The main difference between them is that the HCNN uses the same set of weights for all states, whereas the HSMLP uses a different set of weights for different states. Due to their similarity, the same training and recognition methods for HCNN can be applied to the HSMLP.

V. SIMULATIONS

The effectiveness of these tone recognition schemes discussed above was examined by simulations. Two databases were used in the following experiments. One is for the training and the other for the testing. These two databases comprise well-designed, phonetic-balanced declarative sentential utterances and are composed of almost all 408 types of base syllables of Mandarin speech. The number of syllables in an utterance ranges from 6 to 19. There are, in total, 1268 utterances in the training database. They were uttered by 45 speakers including 30 male and 15 female. The database for testing consists of 189 utterances generated by the same 45 speakers. All utterances were spoken naturally. The speaking rate of an utterance ranges from 2.5 to 4.5 syllables per second. The distribution of five tones in these two databases are shown in Table I. We found that Tone 5 is less frequently appeared in both databases.

First, the basic tone recognition scheme using features extracted only from the currently processing syllable was tested. Following the same method used in [7], ten recognition features were extracted for each syllable. By using an MLP with single hidden layer, the best recognition rate is 80% for the inside test and 76% for the outside test. Comparing with the isolated-syllable case discussed in [7], the performance degrades about 15% for the inside test and 17% for the outside test. Therefore, tone recognition for continuous Mandarin speech is not a trivial problem.

Then, the scheme to compensate the coarticulation effect using additional features extracted from neighboring syllables was examined. A total of 22 input features including ten from the processing syllable, ten from context, and two indicators were used. A recognition rate of 82% was achieved for the outside test. It is much better than that of the basic scheme. By error analysis, we found that the recognition rate for Tone 3 in the experiment is very low when it is followed by another Tone 3. This mainly resulted from the well-known sandhi rule for tone pair 3-3. The sandhi rule indicates that when a syllable of Tone 3 precedes another Tone 3, it will be pronounced approximately as Tone 2. It is therefore no longer distinguishable from Tone 2 by

TABLE II
RECOGNITION RATES OF THE SCHEMES USING CONTEXT FEATURES (SCHEME 1) AND NEIGHBORING TONES (SCHEME 2)
(unit : %)

no. of hidden neurons		60	70	80	90
Scheme 1	inside test	91.39	91.61	92.14	
	outside test	84.21	84.16	83.87	
Scheme 2	inside test		93.16	93.28	93.81
	outside test		84.31	84.80	84.75

simply using acoustic information. To prevent the MLP recognizer from being polluted by these distorted patterns of Tone 3, the tonality of some syllables in these two databases was changed from Tone 3 to Tone 2 in advance. Specifically, all tone bigrams of 3-3 were manually labeled as 2-3, and all tone trigrams of 3-3-3 were labeled as 3-2-3. About 3 and 3.8% syllables were changed from Tone 3 to Tone 2 in the training and the testing databases, respectively. After making the tone change, experimental results by the same recognition scheme are shown in the second and third rows of Table II. A recognition rate of 84.16% was achieved when 70 hidden units were used. By comparing with the previous experiment, we found that not only the recognition rate for Tone 3 was improved significantly, but those for Tone 2 and Tone 5 were improved as well. Due to its effectiveness, the tone modification was applied for all the following experiments.

Then, the scheme to compensate for the effect of sandhi rules by adding tones of two nearest neighboring syllables was tested. Ten more input features were added for indicating tones of both the preceding and the following syllables. A dynamic programming procedure was employed to determine the tone sequence for all syllables in the input utterance based on the criterion of minimal accumulated risk. Experimental results are shown in the fourth and fifth rows of Table II. A recognition rate of 84.80% was achieved when 80 hidden units were used.

Finally, the scheme to compensate the intonation pattern was tested. In both cases using HCNN and HSMLP, the intonation pattern was modeled by a five-state Markov chain with one state for the beginning syllable, one state for the ending syllable, and three states for all intermediate syllables. Experimental results are displayed in Table III. Recognition rates of 85.78 and 86.72% were achieved for the cases of using HCNN and HSMLP, respectively. The performance is better than that obtained in the previous experiment. The effectiveness of the scheme is further analyzed by examining the recognition rates of syllables located in different parts of utterances. Table IV lists the recognition rates of inside test for syllables in three uniformly partitioned parts of utterances. It can be seen from the table that the recognition rates for both the second and the third parts of utterances were improved when the scheme using HCNN or HSMLP was used. This confirmed that the scheme can partially compensate the effect of intonation pattern. We finally check the recognition rates for five tones. Table V lists the recognition rates of five tones for the scheme using HSMLP. It is found that the performances for Tones 1, 2, and 4 are very good, but the recognition rates for Tones 3 and 5 are still far below the average. This mainly results from the relatively high variabilities of F0 contour patterns for both Tones 3 and 5.

VI. CONCLUSIONS

Several tone recognition schemes based on artificial neural networks have been discussed in this correspodence. Both the coarticulation from neighboring context and the intonation pattern of sentence pronunciation had been considered. Experimental results have confirmed that these two effects can be properly compensated for. A recognition rate of 86.72% was achieved.

One possible way to further improve the performance of the system is by incorporating more linguistic information, such as phrasal information and syntactical information, to the system. However, this cannot be independently accomplished without considering the recognition of 408 Mandarin base syllables. Therefore, we will further study this problem in future research on continuous Mandarin speech recognition that integrates tone and base-syllable recognition.

TABLE III
RECOGNITION RATES OF THE SCHEME USING HCNN AND HSMLP
(unit : %)

no. of hidden neurons		65	70	75	80	90
HCNN	inside test		92.26		92.30	92.34
	outside test		85.26		84.70	85.78
HSMLP	inside test	90.92	90.75	90.85		
	outside test	85.78	86.72	86.62		

TABLE IV
RECOGNITION RATES OF INSIDE TEST FOR SYLLABLES IN THREE UNIFORMLY PARTITIONED PARTS OF UTTERANCES
(unit : %)

	hidden units number	position of syllable in the sentence		
		first 1/3	middle 1/3	last 1/3
Scheme 2	80	88.86	86.03	80.03
HCNN	90	89.10	86.54	82.29
HSMLP	70	89.10	87.16	84.38

TABLE V
CONFUSION TABLE FOR THE SCHEME USING HSMLP WITH 70 HIDDEN UNITS

input tone	Recognition result (%)				
	Tone 1	Tone 2	Tone 3	Tone 4	Tone 5
1	88.0	4.5	0.7	5.9	0.9
2	3.3	89.9	4.2	0.9	1.6
3	1.0	13.6	76.6	2.3	6.3
4	4.3	1.5	2.9	90.1	1.1
5	2.1	9.3	11.3	6.2	71.1

REFERENCES

[1] Y. R. Chao, *A Grammar of Spoken Chinese*. Berkeley, CA: Univ. of California Press, 1968; also in *Proc. 1990 IEEE Conf. Acoustic, Speech, Signal Processing*, pp. 517-520.

[2] L. S. Lee, C. Y. Tseng, and M. O.-Young, "The synthesis rules in a chinese text-to-speech system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1309-1320, Sept. 1989.

[3] H. Ma, "Chinese four tone recognition based on the model for process of generating F0 contours of sentences," in *1987 IEEE Conf. Acoust., Speech, Signal Processing*, pp. 65-68.

[4] C. F. Wang, H. Fujisaki, and K. Hirose, "The four tones recognition of continuous chinese speech," in *1990 Int. Conf. Spoken Language Processing*, pp. 221-224.

[5] Y. R. Wang, J.-M. Shieh, and S. H. Chen, "Tone recognition of continuous chinese speech based on hidden Markov model," *Int. J. Patt. Recogn. Artifical Intell.*, vol. 8, no. 1, pp. 233-246, 1994.

[6] J. D. Markel and A. H. Gray, Jr., "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[7] P.-C. Chang, S.-W. Sun, and S.-H. Chen, "Mandarin tone recognition by multilayer perceptron," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[8] J. L. McClelland and D. E. Rumelhart, *Parallel Distributed Processing, Explorations in the Microstructure of Cognition. Vol. 1 : Foundations.* Cambridge, MA: MIT Press, 1986, ch. 9.

[9] E. Levin, "Word recognition using hidden control neural architecture," in *Proc. 1990 IEEE Conf. Acoust., Speech, Signal Processing*, pp. 433-436.

[10] B. P. J. Tebelskis, "Context-dependent hidden control neural network architecture for speech recognition," in *Proc. 1992 IEEE Conf. Acoust., Speech, Signal Processing*, pp. 397-400.