Using GMM for Voiced/Voiceless Segmentation and Tone Decision in Mandarin Continuous Speech Recognition

Ching X. Xu

Northwestern University, Evanston, IL 60208, USA

ABSTRACT

In this paper, methods of Gaussian Mixture Model (GMM) are presented for both silence/voiced/voiceless segmentation and tone decision in Mandarin continuous speech recognition system. GMM has been used for silence/voiced/voiceless segmentation before, but the feature parameters can be modified to improve both accuracy and speed. As a popular method in pattern recognition, GMM is first proposed for tone decision. The two GMMs used are proved to be capable and potential.

1. INTRODUCTION

Voiced/voiceless segmentation and pitch detection are two basic analysis procedures in speech signal processing. They are closely related to each other and both are indispensable in many practical systems. Theoretically, voiced sounds are produced by the periodic vibration of vocal folds and are associated with positive pitch values, while voiceless sounds are produced as irregular noise and are associated with zero pitch value. It is therefore natural to combine voiced/voiceless segmentation and pitch detection. There are many algorithms exploring voiced/voiceless segmentation as a by-product of pitch detection[1-5]. It seems that the combination would greatly improve the processing efficiency. However, the actual results are usually increase of system complication and decrease of system function. Especially at the boundary of voiced and voiceless speech, those algorithms will be in a dilemma to correctly classify the sound. The difficulty originates from the different requirements of these two tasks. Pitch detection would demand relatively long analysis frame. For example, the autocorrelation algorithm needs at least two pitch periods in a frame. On the other hand, voiced/voiceless segmentation would demand relatively short analysis frame. A long frame at the boundary is likely to cover both voiced and voiceless sounds, and will be difficult to be classified. The contradiction in terms of analysis frame becomes more serious in continuous speech recognition because short initials such as b, d and g would probably be concealed by adjacent vowels in a long processing window. Also, it is well known that voiced sounds are only quasi-periodic. Factors during speech production, as well as disturbance from outside, can result in irregular waveforms. If the segmentation between voiced and voiceless sounds depends only on the degree of speech periodicity, the recognition rate will then be greatly decreased. Therefore, we propose to separate voiced/voiceless segmentation from pitch detection. This way, we may be able to use different frame length for these two procedures and introduce more decision features in order to improve the reliability of processing results.

It has been noticed that, although parameters of speech signals are variable and fairly random at certain times, their distributions are on the whole pretty regular. People may have difficulty in getting the real picture of the signals mainly because it is hard to get appropriately involved from the view of subjective experience. As a result, an objective probability statistical model is potentially applicable. Gaussian Mixture Model (GMM) is a simple and efficient approach of pattern recognition, which can smoothly describe the innate distribution of acoustical classification. In this paper, we present it for both silence/voiced/voiceless segmentation and tone decision in Mandarin continuous speech recognition system.

2. SILENCE/VOICED/VOICELESS SEGMENTATION

In Mandarin, most syllables have voiceless onset and all syllables have voiced offset. With silence/voiced/voiceless segmentation, most divisions between and within syllables in Mandarin continuous speech will be accomplished. A complex problem in continuous speech recognition will consequently consist of three relatively easier tasks: classification of initials, recognition of finals and combination of initials with finals. This way, not only recognition techniques can be simplified and storage space saved, but also many conventional skills in isolated speech recognition may be applicable. Furthermore, the probability of combination between initials and finals will be able to improve the recognition of initials, which usually have less conspicuous features than finals and are more difficult to recognize.

There have been many algorithms based on classification of certain characteristic parameters in voiced/voiceless segmentation[6-10]. Although most of them are successful to some extent, there is still much space for development in this area. The first crucial problem is the selection of characteristic parameters. Surely we would like to use the ones which are easier to detect and can validly classify the three status of speech. According to experience, we take the following six parameters into consideration.

(1) Logarithm of summation of absolute value of amplitude — A_n

$$A_n = 10 \times \log_{10}[e + \sum_{n=0}^{N_1 - 1} |S(n)|] \quad (1)$$

where ε is a small positive value to avoid any illegal calculation if its following item equals zero, and is assigned as 10^{-9}; N_1 is the length of analysis window, and is assigned as a short frame with 64 points; S(n) is the digital speech signal. It is clear that A_n is also half of the logarithm energy.

(2) Corrected zero-crossing rate — $Z_c^{'}$. Zero-crossing rate Z_c is an important parameter in speech signal. Under ideal conditions, voiceless sound should associate with high Z_c, voiced sound with low Z_c, and silence with 0 Z_c. However, the existence of noise in real condition will bias the actual value of Z_c. Consequently, we employ corrected zero-crossing rate $Z_c^{'}$ to replace Z_c. First, we identify the noise level of speech signal according to the starting and end portions of the signal. Then we use the noise level as the zero point and calculate the corresponding zero-crossing rate.

(3) Auto-correlation function of adjacent speech samples — R_1.

$$R_1 = \frac{1}{N_1} \sum_{n=0}^{N_1-2} S(n)S(n+1) \qquad (2)$$

where N_1 is the length of analysis window, and S(n) is the speech signal, as in (1).

(4) The first coefficient of 12-rank LPC — α_1[11].

(5) Normalized LPC prediction difference energy — P

$$P \equiv E_s - E_p \qquad (3)$$

where $$E_s = 10 \times \log_{10}[e + \frac{1}{N_2} \times \sum_{n=0}^{N_2-1} S^2(n)]$$

$$E_p = 10 \times \log_{10}\{e + \left| \frac{1}{N_2} \times \sum_{n=0}^{N_2-1} S^2(n) - \sum_{n=1}^{M} [a(n) \times r(n)] \right| \}$$

and $$r(n) = \frac{1}{N_2} \times \sum_{k=0}^{N_2-n-1} S(n)S(k+n)$$

where N_2 is the length of analysis window, which is assigned as a long frame with 256 points to improve the frequency resolution; and M is the rank of LPC predictor, i.e. 12. Other variables are the same as in (1).

(6) High/low frequency energy ratio — k. The energy of voiced speech mainly distributes around 1kHz and below, while the energy of voiceless speech mainly distributes around 2kHz and above. As a result, the value of k may help us to differentiate voiceless sound from voiced sound. We use 1.5kHz as the boundary and calculate the ratio of high frequency energy versus low frequency energy. High frequency energy refers to the energy between 1.5kHz and 6kHz, and low frequency energy refers to the energy between 0 and 1.5kHz.

A GMM with these six parameters is implemented and the result is positive. The data materials we use are part of the continuous speech database of Chinese National 863 Project. There are totally 1560 sentences, which consist of three groups: Group A with 521 sentences, Group B with 519 sentences and Group C with 520 sentences. The subjects for each group are different, but are in similar distribution: half males, half females, and age from 16 to 60 years old. Their speech productions are recorded in a relatively quiet laboratory environment, with 16kHz sampling rate and 16 bits accuracy. We randomly choose one subject from each group and name them as A, B and C. Thirty sentences from A and thirty sentences from B are used as training data. These sixty sentences cover all initials and finals. Other forty-four sentences of A and nine sentences of C are used as testing data. The fifty-three sentences not only cover all initials and finals, but also guarantee that each sound appears in the testing data at least four times. All training and testing sentences are manually labeled with starting and end points of each initial and final, as well as corresponding Pinyin, in order to provide reliable references. The recognition rates for voiced and voiceless segments are 97% and 84.9%, respectively. However, such a system will not be able to run on-line because of the large amount of calculation. In order to speed up the processing, and retain or even further improve the segmentation accuracy, we explore the correlations of the six parameters. The three models in this six-dimension GMM are summarized as following:

Silence: $\begin{vmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 \end{vmatrix}$ Voiced: $\begin{vmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{vmatrix}$ Voiceless: $\begin{vmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{vmatrix}$

Correspondingly, the correlations of each parameter with other parameters in individual models are graded as following, and the total grades of all parameters (including the correlations with themselves) are also shown:

	Silence	Voiced	Voiceless	Total
A_n	5	1	2	8
Z_c'	2	2	5	9
R_1	4	4	5	13
α_1	5	1	5	11
P	5	3	4	12
k	5	3	5	13

The higher grade means the larger correlation of a parameter with other parameters and the less contribution of this parameter in a multi-dimension GMM. A_n and Z_c' are meaningful and simple, and also have the lowest correlation grades. They are the first choices for an efficient GMM. Nevertheless, the GMM with only these two feature parameters is not practical enough. Its recognition rates for three kinds of segments are: silence 93.4%, voiced 95.0% and voiceless 68.2%. A third dimension needs to be added to the model. Although α_1 has the third lowest correlation grade, the grade related to A_n and Z_c' of it is the same as that of R_1. Since α_1 and R_1 are consistent in terms of physical meaning, and α_1 needs complicated calculation in frequency domain while R_1 is much easier to get, we choose R_1 rather than α_1 as the third feature parameter.

A GMM with A_n, Z_c' and R_1 as the measured parameters is implemented along with a good five-dimension GMM[7]. They both get very promising results. The recognition rates for testing data are shown in Table 1.

Type of Sound	Number of Frames	Recognition Rate		Number of Segments	Recognition Rate	
		5-D GMM	3-D GMM		5-D GMM	3-D GMM
Silence	24888	95.4%	96.3%	360	78.9%	87.2%
Voiced	24585	96.0%	94.8%	1351	98.4%	99.0%
Voiceless	7846	74.8%	92.1%	433	83.4%	97.0%

Table 1 Recognition rates of two GMMs

The relatively low accuracy for silence segment will not have a big effect except that it needs further classification in post processing. During the experiments, we also notice that most mistakes take place around the boundary of different speech segments and mainly result in shortening or lengthening the corresponding segments. Since the duration of a sound in Mandarin is generally not communicative, as it is in English and some other languages, and the analysis frames with conspicuous features have been retained, the final recognition results of the system regarding continuous speech will not be influenced much. Generally speaking, the five-dimension GMM has good performance in terms of segmentation, while the three-dimension GMM does even better. The most important improvement is that the accuracy for recognizing voiceless sound is increased by 17.3% in terms of frame, and 13.6% in terms of segment. Since consonants in syllables usually carry more speech information than vowels do because of their explicit articulation places and manners, the right classification of voiceless sound will not only decrease the complexity of post processing, but also greatly improve final syllable recognition. At the same time, the three-dimension GMM uses relatively simple parameters so that it improves about 50% of the system speed than the five-dimension GMM. Also, the method is applicable to speaker-independent system because it has high accuracy for open data — the speech of C.

3. TONE DECISION

Tone information is crucial in Mandarin. Its application in recognition will greatly reduce the ambiguity and complicacy of language understanding. Moreover, the association possibility of initials, finals and tones may be a reference for checking the reliability of individual recognition. It is hard, however, to correctly classify tones in continuous speech. The tonal status in continuous speech is much more complicated than it in isolated speech. Since almost 75% of Mandarin words are disyllabic words, and words with three or more syllables can be regarded as combinations of disyllabic and/or monosyllabic words in terms of structure [12], phoneticians often concentrate on disyllabic words when they intend to study tonal variations. Furthermore, the flow of Mandarin continuous speech is sort of a sequence of words[13]. Therefore, we assume that the tone status in disyllabic words covers most of the conditions in continuous speech. Neutral Tone is excluded to simplify the problem.

Since differences among tones are robust except tone sandhi, we propose to use GMM for tone decision, as well as for silence/voiced/voiceless segmentation. After segmentation, the voiced segments are further segmented into voiced initials and finals [14]. The fundamental frequency (F_0) of finals is extracted by using FAD[15]. It has been noticed that the starting 60ms and ending 40-50ms signals are not really related to the tone decision[11], so 960 sample points at the beginning and 640-800 sample points at the end are omitted. The F_0 curve of remaining middle points are used for tone decision for each final. Five points with almost equal distance from adjacent ones on each F_0 contour are selected as T_1, T_2, T_3, T_4 and T_5. The four normalized differences between every two adjacent points

$$\frac{T_i - T_{i+1}}{T_3} (i = 1,2,3,4)$$

are calculated as the feature parameters for GMM. The normalization has two advantages. One is to differentiate different tone values with similar F_0 variation trend, and the other is to adapt to different speakers.

According to previous studies [12, 16-18], there are several possible tone values for each tone in disyllabic words.

High Tone: 55, 54 and 44;
Rising Tone: 35, 34 and 55;
Low Tone: 214, 21, 22, 34, 35 and 55;
Falling Tone: 51, 53, 41, 52 and 55.

The first value of each tone is the classic value of it. Tones are associated with their classic values in isolated speech. The other values are variations originate from carryover or anticipatory effect in continuous speech. It is clear that some different tone values have similar F_0 variation trend, like 55, 44 and 22. Since the measured parameters in GMM are normalized, these tone statuses can be differentiated from each other without problems. The repetitions of tone values, such as 55 in all four tones, will be able to be correctly classified by language models, which are beyond this paper. We now skip the less frequent repetitions and summarize the 12 possible tone values as following.

High Tone: 55, 54 and 44;
Rising Tone: 35 and 34;
Low Tone: 214, 21 and 22;
Falling Tone: 51, 53, 41 and 52.

Correspondingly, we train 12 models in GMM.

The experimental data are speech production of a female subject, under a relatively quiet laboratory environment, with 10kHz sampling rate and 8 bits accuracy. There are 1272 final segments from the continuous speech. We use 144 segments, which include above 12 tone values and are evenly distributed, as the training data, and the total 1272 segments as the testing data. The recognition rate is shown in the last column of Table 2. If we take only the classic value for each tone into consideration, i.e., train four models in GMM, the results will be much worse, as also shown in Table 2.

Type of Tones	Number of Segments	Recognition Rate	
		4 Models	12 Models
High Tone	335	45.3%	84.3%
Rising Tone	260	71.2%	75.4%
Low Tone	323	59.4%	81.7%
Falling Tone	354	74.6%	81.9%

Table 2 Tone recognition rates of a GMM with two different numbers of models

The comparison in Table 2 further provides evidence for more tone status in continuous speech than in isolated speech. Only with more detailed description of tone conditions, can we perform better in tone decision. On the other hand, the results of 12-model GMM are still not ideal. Along with the observation of various tone values in actual speech, it implies the necessity of further exploration in phonetic facts. Also, the expendability to speaker-independent system of this method will need more testing and research.

4. ACKNOWLEDGEMENT

The work presented in this paper was done as part of the author's master thesis [19] two years ago in Institute of Acoustics, Academia Sinica. The author appreciates the direction and help of Professor Tiecheng Yu.

5. REFERENCES

1. Lahat, M. "A spectral autocorrelation method for measurement of the fundamental frequency of noise — corrupted speech", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-35, No. 6, pp. 741-750, June, 1987.*

2. Markel, J. D. "The SIFT algorithm for fundamental frequency estimation", *IEEE Trans. Audio Electroacoust., Vol. AU-20, No. 5, pp. 367-377, December, 1972.*

3. Miller, N. J. "Pitch detection by data reduction", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-23, No. 1, pp. 72-79, February, 1975.*

4. Noll, A. M. "Cepstrum pitch determination", *J. Acoust. Soc. Amer., Vol. 41, pp. 293-309, February, 1967.*

5. Ross, M. J. et al. "Average magnitude difference function pitch extractor", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-22, No. 5, pp. 353-362, October, 1974.*

6. Atal, B. S. & Hanauer, S. L. "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer., Vol. 50, pp. 637-655, August, 1971.*

7. Atal, B. S. & Rabiner, L. R. "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, No. 3, pp. 201-212, June, 1976.*

8. Schroeder, M. R. "Vocoders: analysis and synthesis of speech", *Proc. IEEE, Vol. 54, pp. 720-734, May, 1966.*

9. Siegel, L. J. "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, No. 1, pp. 83-89, February, 1979.*

10. Siegel, L. J. & Bessey, A. C. "Voiced/unvoiced/mixed excitation classification of speech", *IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-30, No. 3, pp. 451-460, June, 1982.*

11. Yang, X. & Chi, H. *Digital processing of speech signal (in Chinese)*, Electronics Industry Press, 1995.

12. Wu, Z. *Essentials of modern Chinese phonetics (in Chinese)*, Foreign Language Printing House, 1992.

13. Wu, Z. "Tonal variations in Mandarin sentences (in Chinese)", *Chinese Linguistics, 439-450, 1982.*

14. Liu, J. *Chinese continuous speech recognition and algorithm research in voiced speech segmentation (in Chinese)*, Master thesis, Institute of Acoustics, Academia Sinica, 1996.

15. Xu, X. *A new method exploration in pitch detection (in Chinese)*, Bachelor thesis, Peking University, 1995.

16. Chen, Y. & Wang, R. *Language Signal Processing (in Chinese)*, Chinese Science & Technology University Press, 1990.

17. Huang, B. & Liao, X. *Modern Chinese (in Chinese)*, Gansu People's Press, 1983.

18. Li, Z. "The illustration analysis of Chinese disyllabic tones and its application in synthetic language (in Chinese)", *Journal of Acoustics, Vol. 10, No. 2, pp. 73-84, 1985.*

19. Xu, X. *Research on problems in Chinese continuous speech recognition — silence/voiceless/voiced segmentation and tone recognition (in Chinese)*, Master thesis, Institute of Acoustics, Academia Sinica, 1998.