International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:5, 2011

Recognition by Online Modeling – a New Approach of Recognizing Voice Signals in Linear Time

Jyh-Da Wei, Hsin-Chen Tsai

Abstract—This work presents a novel means of extracting fixed-length parameters from voice signals, such that words can be recognized in linear time. The power and the zero crossing rate are first calculated segment by segment from a voice signal; by doing so, two feature sequences are generated. We then construct an FIR system across these two sequences. The parameters of this FIR system, used as the input of a multilayer proceptron recognizer, can be derived by recursive LSE (least-square estimation), implying that the complexity of overall process is linear to the signal size. In the second part of this work, we introduce a weighting factor λ to emphasize recent input; therefore, we can further recognize continuous speech signals. Experiments employ the voice signals of numbers, from zero to nine, spoken in Mandarin Chinese. The proposed method is verified to recognize voice signals efficiently and accurately.

Keywords—Speech Recognition, FIR system, Recursive LSE, Multilayer Perceptron

I. Introduction

THE difficult of speech recognition is primarily due to the dynamic speed of speaking words[1], [2]. A voice signal recognizer usually needs to temporally align the features of the test utterance with those of reference utterances [3], [4], [5], causing conventional recognition approaches to be quite time-consuming. Such conventional approaches include Dynamic Time Warping (DTW) [6] and the Hidden Markov Model (HMM) [7], [8].

This work develops a novel method that can recognize voice signals in linear time, which goal requires that fixed-length parameters be extracted from voice signals. To do this, we consider that a voice signal may reflect a system embracing potential input and output information. Extracting short-term features segment by segment, we can get input and output sequences and thus form a system with those fixed-length parameters we need for recognition. In this paper, we select the power and the zero crossing rate as feature sequences, referred to as P and Z respectively. We then construct a Finite Impulse Response (FIR) [9] system which transfers input P to output S. The parameters of this FIR system finally serve as the input of a multilayer proceptron recognizer.

The rest of this paper is organized as follows. In Section II, we explain the recognition process, including how to use recursive least-square estimation (LSE) [10], [11] for system modeling. In Section III, we present a method for continuous speech recognition, which needs a weighting factor λ included to emphasize recent input. Conclusion and feature work are finally described in Section IV.

Jyh-Da Wei, the communicating author, is with the Department of Computer Science and Information Engineering, Chang-Gung University, Taoyuan, Taiwan.

Hsin-Chen Tsai is with the Department of Computer Science and Information Engineering, Chang-Gung University, Taoyuan, Taiwan.

II. Recognition by Online Modeling

Figure 1 shows the framework of our method. Initially, a voice signal with one word to recognize is recorded (or already separated from a continuous voice signal). A sequence, V, is generated after normalizing the amplitude into [-1,1]. Next, the power (P) and the zero-crossing rate (Z) sequences are derived, according to Eqs. (1) and (2).

$$P[k] = \frac{1}{N} \sum_{n=W\cdot k}^{W\cdot k+N-1} (V[n])^2, \tag{1}$$

$$Z[k] = \frac{1}{2N} \sum_{n=W\cdot k}^{W\cdot k+N-1} |sig(V[n+1]) - sig(V[n])|, \tag{2}$$

where N is the window size, W is the sliding distance between two windows, and

$$sig(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & otherwise. \end{cases}$$

Regarding the voice utterances as systems with $I - O$ sequences of P and Z, this work then constructs an FIR (Finite Impulse Response) system corresponding to the recorded voice signal. An FIR system is described by their parameters. Therefore, a multiplayer perceptron (MLP) is finally used to map these parameters onto a reference word, by assuming that voice signals of the same word are associated with resembling FIR systems.

An online (recursive) algorithm of the LSE (Least Squares Estimation) can be used to determine the parameters of FIR systems. Accordingly, this voice recognition method is called, "Recognition by Online Modeling". Using fourth-order moving average as an example, the recursive LSE method is described as follows:

$$Q_k = Q_{k-1} - \frac{Q_{k-1}\phi_k\phi_k^T Q_{k-1}}{1 + \phi_k^T Q_{k-1}\phi_k}, \tag{3}$$

$$\theta_k = \theta_{k-1} + Q_k\phi_k(Z_k - \phi_k^T\theta_{k-1}), \tag{4}$$

where Q_0 is a diagonal matrix with 5 rows of sufficiently large entry values, θ is a zero vector with 5 entries, and

$$\phi_k = [\ P[k]\ P[k-1]\ P[k-2]\ P[k-3]\ 1]^T.$$

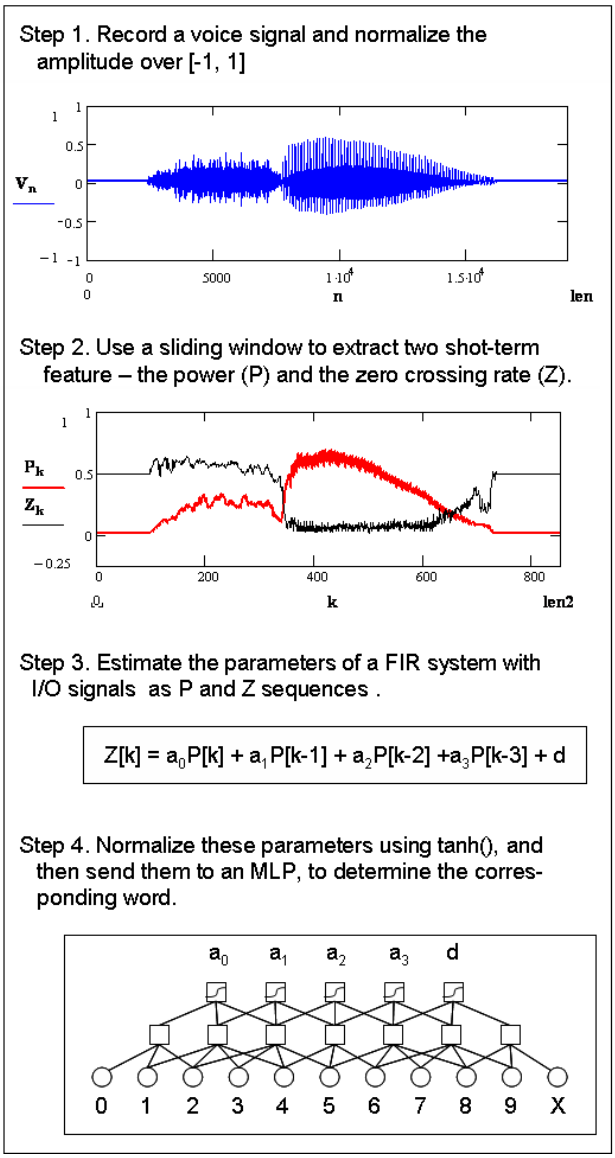

Fig. 1. *Recognition by Online Modeling.* This method involves four steps, recognizing voice signals in linear time.

The parameters of our FIR system,i.e., a_0, a_1, a_2, a_3 and d, will appear in the last θ vector.

The experiments of this paper considered voice signals of numbers zero to nine, spoken in Mandarin Chinese. These signals were recorded using a mono channel at a 22,050 Hz sampling rate and 8-bit quantization. The P and Z feature sequences were extracted using a window of size N=220 and a window sliding distance of W=22. An FIR systems of a fourth-order moving average were subsequently constructed with $I-O$ set to $P-Z$. The parameters were mapped onto reference words by an MLP that used one hidden layer of 20 nodes, trained at a learning rate of 0.05 and 2000 training epochs.

This work involved twenty independent experiments to verify the proposed method. Each experiment adopted voice signals spoken by one person. The speaker spoke each word 30 times – 20 for training and the other 10 for testing. Table 1 lists the average accuracy of these experiments. This table indicates that 97.5% words could be recognized, referring to the highest three output nodes of the MLP. This finding implies that the parameters extracted by the proposed method can be used as recognition features.

III. Continuous Speech Recognition

The proposed method can be extended to recognize continuous speech signals. For this purpose, we use LSE with weighing factor λ to emphasize recent input signals. The online estimation equations are now as follows.

$$Q_k = \frac{1}{\lambda}[Q_{k-1} - \frac{Q_{k-1}\phi_k\phi_k^T Q_{k-1}}{\lambda + \phi_k^T Q_{k-1}\phi_k}], \tag{5}$$

$$\theta_k = \theta_{k-1} + Q_k\phi_k(Z_k - \phi_k^T\theta_{k-1}), \tag{6}$$

Figure 2(a) is a voice signal which records words “5”, “6” and “7” spoken continuously. Figure 2(b) is the feature sequences P and Z corresponding to Fig. 2(a). As shown in Fig. 2(c)(d), the influence of input data in weighted LSE is getting reduced along time passing. Figure 3 shows the parameter sequences extracted online by weighted LSE. Target output words corresponding to some recognizable segments included “5”, “6” and “7”. Except from these regions, we marked the output as ”X” to indicate an unrecognizable situation.

Weighted LSE also enables the recognizer to automatically divide the words out of continuous speech signals. Here we introduce two thresholds T_p and T_n. Once the power value is less than T_p and the parameters are unrecognizable, we sign a separate stamp on the time series. Between two separate stamps, there is at most one word that can be recognized. We count and get the most frequently recognized word in this interval. If the count is not less than T_n, we output this word; otherwise, none of words will be generated within this interval and we continue to detect the next separate stamp.

We repeat the above experiments except that the testers continuously speak eight numbers at random. Table 2 shows the experimental results with λ, T_p, T_n set to $0.96,\ 0.1$ and 5 respectively. Continuous speech recognition is more difficult then single word recognition. As Table 2 shown, the accuracy cannot be higher than 90% for testing data.

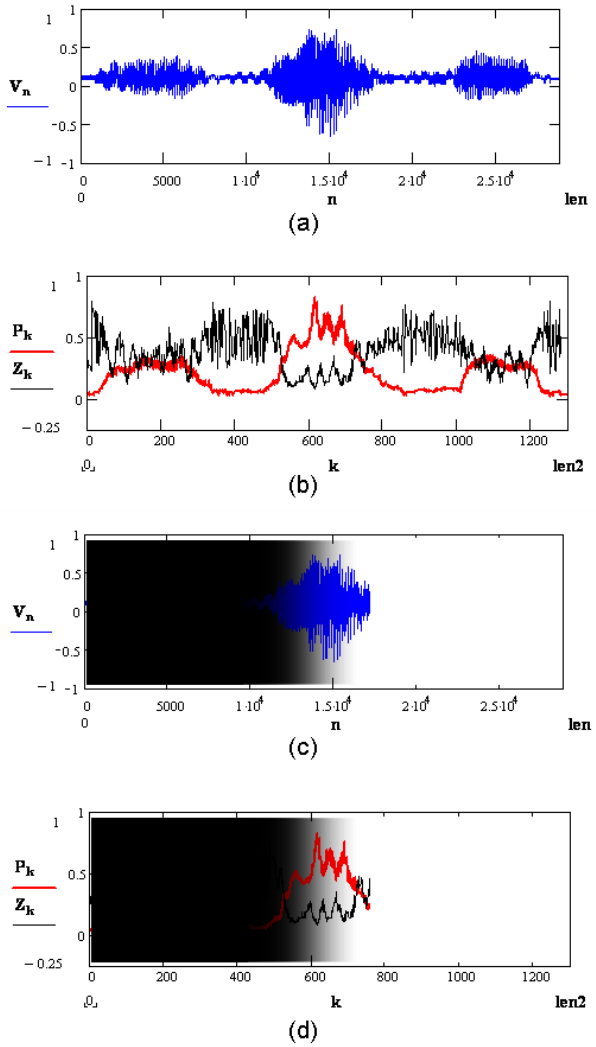

Fig. 2. *Weighted LSE for continuous speech recognition.* Recursive Least-square Estimation method with a weighting factor λ can be used to recognize recent input in a continuous speech signal.

IV. Conclusion

This work develops a novel means of extracting fixed-length parameters from voice signals that are spoken at dynamic speed. Accordingly, reference words can be recognized in linear time. Experimental results confirm that the proposed method provides an efficient speech recognizer.

Additionally, the proposed method is especially applicable to hardware design. Engineers can implement the parameter extractor and the MLP in parallel, reducing the complexity of circuitry and recognizing voice signals almost in real time. Consequently, recognition by online modeling is a new direction for speech recognition, which merits further attention.

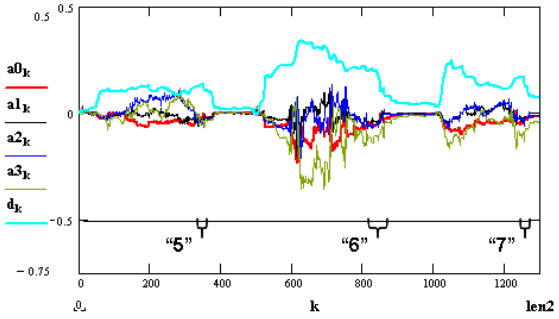

Fig. 3. *Parameters extracted by weighted LSE.* Five parameter sequences $a0$, $a1$, $a2$, $a3$ and d were extracted online by weighted LSE. Target output words corresponding to some recognizable segments included "5", "6" and "7". Except from these regions, we marked the output as "X" to indicate an unrecognizable situation.

TABLE I
Average accuracy of 20 speaker-dependent experiments

Referring to	Training results	Testing results
Highest output	98.0 %	84.5 %
Highest 2 outputs	99.5 %	91.0 %
Highest 3 outputs	100.0 %	97.5 %

TABLE II
Average accuracy of 20 speaker-dependent experiments with continuous speech signals

Referring to	Training results	Testing results
Highest output	94.2 %	79.6 %
Highest 2 outputs	95.6 %	84.2 %
Highest 3 outputs	98.4 %	89.5 %

Acknowledgment

This work was supported in part by the Republic of China National Science Council (grant no. NSC98-2218-E-182-005).

References

[1] X. Huang and A. Acero and H. W. Won, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, NJ, 2001.
[2] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, McGraw-Hill, New York, NY, 2001.
[3] J.R.Deller and J.G.Proakis and J. H. L. Hansen, *Discrete Time Processing of Speech Signals*, Mac Millan, 1993.
[4] Tanja Schultz, Alan W. Black, Stephan Vogel, and Monika Woszczyna, "Flexible speech translation systems," *IEEE Transactions on Audio, Speech, and Language Processin*, vol. 14, pp. 403–411, 2006.
[5] William M. Campbell, Joseph P. Campbell, Douglas A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 210–229, 2006.
[6] N.U. Nair and T.V. Sreenivas, "Multi pattern dynamic time warping for automatic speech recognition," in *Proc. 2008 IEEE Region 10 Conference*, 2008, pp. 1–6.
[7] Peter Jančovič and Münevver Köküer, "Incorporating the voicing information into hmm-based automatic speech recognition in noisy environments," *Speech Commun.*, vol. 51, no. 5, pp. 438–451, 2009.

[8] Ángel de la Torre, Antonio M. Peinado, Antonio J. Rubio, José C. Segura, and Carmen Benítez, "Discriminative feature weighting for hmm-based continuous speech recognizers," *Speech Commun.*, vol. 38, no. 3-4, pp. 267–286, 2002.
[9] Alan V. Oppenheim and Ronald W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 2009.
[10] Ben M. Chen and Kemao Peng and Tong H. Lee and Venkatakrishnan Venkataramanan, *System Modeling and Identification*, Springer London, 2006.
[11] J. P. Norton, *An Introduction to Identification*, Dover Publications, Inc., New York, NY, USA, 2009.