論文 / 著書情報
Article / Book Information

Title	Noise Robust Speech Recognition Using F0 Contour Extracted by Hough Transform
Author	Koji Iwano, Takahiro Seki, Sadaoki Furui
Journal/Book name	7th International Conference on Spoken Language Processing (ICSLP-2002), Vol. , No. , pp. 941-944
発行日 / Issue date	2002, 9

NOISE ROBUST SPEECH RECOGNITION USING F_0 CONTOUR EXTRACTED BY HOUGH TRANSFORM

Koji Iwano, Takahiro Seki, and Sadaoki Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
{iwano,tseki,furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes a noise robust speech recognition method using prosodic information. In Japanese, fundamental frequency (F_0) contour represents phrase intonation and word accent information. Consequently, it conveys information about prosodic phrase and word boundaries. This paper first proposes a noise robust F_0 extraction method using Hough transform, which achieves high extraction rates under various noise environments. Then it proposes a robust speech recognition method using syllable HMMs which model both segmental spectral features and F_0 contours. Speaker-independent experiments are conducted using connected digits uttered by 11 male speakers in various kinds of noise and SNR conditions. The recognition accuracy is improved in all noise conditions, and the best absolute improvement of digit accuracy is about 4.7%. This improvement is achieved due to the more precise digit boundary detection by the robust prosodic information.

1. INTRODUCTION

Recently, continuous speech recognition has made great progress and high recognition rates can be achieved for read speech uttered in a clean/quiet environment. However, speech recognition technology is not yet able to provide adequate performance in spontaneous speech tasks or when used in a noisy environment.

Given the importance of prosodic features, such as word accent and phrase intonation, in the human speech perception process, several experiments using prosodic features in machine speech recognition process have been conducted[1]. However, prosodic information is hardly utilized in current speech recognition systems.

Spontaneous speech includes a lot of repairs and disfluencies, and they decrease the recognition performance[2]. It has been reported that the performance of spontaneous speech recognition can be improved by predicting the appearance of these events using prosodic features such as pause length and syllable duration[3].

Although various methods have been proposed for noise-robust speech recognition, prosodic information has not yet been used, because of the difficulty of correctly extracting prosodic features in noisy environments. If F_0 values can be extracted robustly, they should be useful for robust speech recognition. Since fundamental frequency (F_0) contours represent phrase intonation and word accent in Japanese utterances, they are expected to be useful to detect prosodic phrases and word boundaries.

From these viewpoints, a robust F_0 extraction method using Hough transform and a robust speech recognition scheme using the F_0 contours are proposed in this paper.

This paper is organized as follows: In Section 2, a robust F_0 extraction method using Hough transform is proposed. Section 3 describes our modeling scheme for noise robust speech recognition using syllable HMMs combining segmental and prosodic information. Experimental results are reported in Section 4, and Section 5 concludes this paper.

2. F_0 EXTRACTION USING HOUGH TRANSFORM

2.1. Hough Transform

Hough transform is a technique to robustly extract parametric patterns, such as lines, circles, and ellipses, from a noisy image[4].

A Hough transform method to extract a significant line from an image on x–y plane can be formulated as follows. Suppose the image consists of n pixels at (x_i, y_i) $(i = 1, \cdots, n)$. Every pixel on the x–y plane is transformed to a line on the m–c plane as

$$c = -x_i m + y_i \quad (i = 1, \cdots, n) \tag{1}$$

Brightness value of the pixel on the x–y plane is accumulated at every point on the line. This process is called "voting" to the m–c plane. After voting of all pixels, the maximum accumulated voting value on the m–c plane is detected, and the peak point (m, c) is transformed to a line on the x–y plane by the following equation:

$$y = mx + c \tag{2}$$

2.2. F_0 Extraction Using Hough Transform

Although F_0 contours have temporal continuity in the voiced period, the cepstral peaks which have been widely used to extract the F_0 values often cause errors, including half pitch, double pitch and drop outs, due to noise effects. To take advantage of the continuity, the Hough transform is applied to time-cepstrum images of noisy speech.

Speech waveforms are sampled at 16kHz and transformed to 256 dimensional cepstra. A 32ms-long Hamming window is used to extract frames every 10ms. To the time-cepstrum image, a nine frame moving window is applied at every frame interval to extract an image for line information detection. The time-cepstrum image is used as the pixel brightness image for the Hough transform. An F_0 value is obtained from a cepstrum index of the center point for the detected line. Since the moving window has nine frames, the time continuity for 90ms is considered in this method.

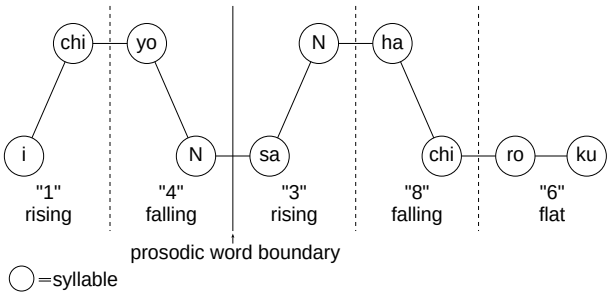


Fig. 1. An example of F_0 contour of Japanese connected digit speech.

In conventional F_0 extraction methods, F_0 values are extracted independently at every frame and various smoothing techniques are applied afterwards. The problem of these methods is that they are sensitive to a decrease in the correctness of the raw F_0 values. Since our method uses the continuity of cepstral images, it is expected to be more robust than conventional methods.

2.3. Evaluation of F_0 Extraction

Utterances from two speakers, one male and one female, were selected from the ATR continuous speech corpus to evaluate the proposed method. Each speaker uttered 50 sentences. This corpus has correct F_0 labels given manually. White noise, in-car noise, exhibition-hall noise, and elevator-hall noise were added to these utterances at three SNR levels: 5, 10, and 20dB. Accordingly, 1,200 utterances were made for evaluation.

The correct F_0 extraction rate was defined as the ratio of the number of frames in which the extracted values were within $\pm 5\%$ from the correct F_0 values to the total number of labeled voice frames.

Evaluation results showed that the extraction rate averaged over all noise conditions was improved by 11.2%, in absolute value from 63.6% to 74.8%, compared to the conventional method without smoothing.

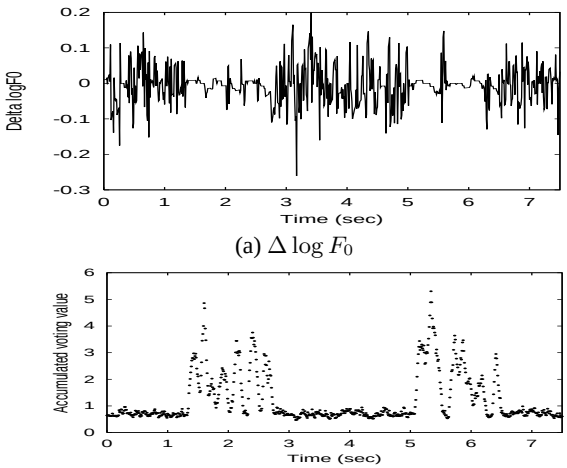
3. INTEGRATION OF SEGMENTAL AND PROSODIC INFORMATION FOR NOISE ROBUST SPEECH RECOGNITION

3.1. Japanese Connected Digit Speech

The effectiveness of the F_0 information extracted by the proposed method on speech recognition was evaluated in a Japanese connected digit speech recognition task. In Japanese connected digit speech, two or three digits often make one prosodic phrase. Figure 1 shows an example of F_0 contour of connected digit speech. The first two digits make the first prosodic phrase, and the latter three digits make the second prosodic phrase. The transition of F_0 is represented by syllabic units, and each syllable can be prosodically labeled as a "rising", "falling", or "flat" F_0 part. Since this F_0 feature changes at digit boundaries, the accuracy of digit alignment in the recognition process is expected to be improved by this information.

3.2. Integration of Segmental and Prosodic Features

Each segmental feature vector has 25 elements consisting of 12 MFCC, their delta, and the delta log energy. The window length is 25ms and the frame interval is 10ms. Cepstral mean subtraction (CMS) is applied to each utterance.



(a) $\Delta \log F_0$

(b) Accumulated voting value

Fig. 2. An example of the prosodic features in Japanese connected digit speech for a male speaker's utterance, "9053308" "3797298", with 20dB SNR white noise.

Prosodic feature vectors have two elements; one is $\Delta \log F_0$ value which represents the F_0 transition, and the other is the accumulated voting value obtained in the Hough transform which indicates the degree of temporal continuity in the F_0.

$\Delta \log F_0$ value is calculated as follows:

$$\Delta \log F_0 = \frac{d \log F_0}{dt} \quad (3)$$

$$= \frac{d \log F_0}{dF_0} \cdot \frac{dF_0}{dt} \quad (4)$$

$$= \frac{1}{F_0} \cdot \Delta F_0 \quad (5)$$

ΔF_0 is directly computed from the line extracted by the Hough transform. An example of the time function of $\Delta \log F_0$ and accumulated voting values is shown in Figure 2. A male speaker's utterance, "9053308" "3797298", with white noise added at 20dB SNR is shown. In unvoiced and pause periods, both $\Delta \log F_0$ and accumulated voting values are more fluctuating than in voiced periods. These features are expected to be effective to detect boundaries between voiced and unvoiced/pause periods.

The segmental and prosodic feature vectors are combined for each frame to build a 27 dimensional segmental-prosodic feature vector.

3.3. Multi-stream Syllable HMMs

3.3.1. Basic Structure of Syllable HMMs

Since syllable variation and the change of F_0 transition such as "rising", "falling" and "flat" are highly related, the segmental and prosodic features are integrated using syllabic unit HMMs. Our preliminarily experiments show that syllable HMMs and tied-state triphone HMMs have approximately the same digit recognition accuracy for the connected digit task.

The integrated syllable HMM denoted by "SP-HMM (Segmental-Prosodic HMM)" is modeled taking the context and the F_0 transition into account. Table 1 is the list of SP-HMMs used in our experiments. Each Japanese digit uttered continuously with other

Table 1. List of SP-HMMs (Segmental-Prosodic HMMs). SP-HMM is denoted by either "LC-SYL,PM" or "SYL+RC,PM". "LC-SYL" indicates the left-context dependent syllable and "SYL+RC" indicates the right-context dependent syllable. "PM" indicates F_0 pattern which is either rising("U"), falling("D"), or flat("F").

digit	model		
0	ze+ro,U	ze+ro,D	ze+ro,F
/zero/	ze-ro,U	ze-ro,D	ze-ro,F
1	i+chi,U	i+chi,D	i+chi,F
/ichi/	i-chi,U	i-chi,D	i-chi,F
2	ni+i,U	ni+i,D	ni+i,F
/ni:/	ni-i,U	ni-i,D	ni-i,F
3	sa+N,U	sa+N,D	sa+N,F
/saN/	sa-N,U	sa-N,D	sa-N,F

digit	model		
4	yo+N,U	yo+N,D	yo+N,F
/yoN/	yo-N,U	yo-N,D	yo-N,F
5	go+o,U	go+o,D	go+o,F
/go:/	go-o,U	go-o,D	go-o,F
6	ro+ku,U	ro+ku,D	ro+ku,F
/roku/	ro-ku,U	ro-ku,D	ro-ku,F
7	na+na,U	na+na,D	na+na,F
/nana/	na-na,U	na-na,D	na-na,F

digit	model		
8	ha+chi,U	ha+chi,D	ha+chi,F
/hachi/	ha-chi,U	ha-chi,D	ha-chi,F
9	kyu+u,U	kyu+u,D	kyu+u,F
/kyu:/	kyu-u,U	kyu-u,D	kyu-u,F
	sil	sp	

digits can be modeled by a concatenation of two syllables. Even "2" (/ni/) and "5" (/go/) can be modeled by two syllables since their final vowel is often lengthen as /ni:/ and /go:/. Context of each syllable is considered only within each digit in our experiment. Therefore, the SP-HMM can be denoted by either a left-context dependent syllable "LC-SYL,PM" or a right-context dependent syllable "SYL+RC,PM", where "PM" indicates the F_0 transition pattern which is either rising ("U"), falling("D") or flat("F"). For example, "the first syllable /i/ of "1" (/ichi/) which has rising F_0 transition" is denoted as "i+chi,U". Each SP-HMM has a standard left-to-right topology with $n \times 3$ states, where n is the number of phonemes in the syllable. The "sil" and "sp" models are used for a silence between digit strings and a short pause between digits, respectively.

3.3.2. Multi-stream Modeling

SP-HMMs are modeled as multi-stream HMMs. In recognition, the probability $b_j(\boldsymbol{O}_{SP})$ of generating segmental-prosodic observation $\boldsymbol{O}_{SP}$ at state j is calculated by:

$$b_j(\boldsymbol{O}_{SP}) = b_j(\boldsymbol{O}_S)^{\lambda_S} \cdot b_j(\boldsymbol{O}_P)^{\lambda_P} \tag{6}$$

where $b_j(\boldsymbol{O}_S)$ is the probability of generating segmental features $\boldsymbol{O}_S$, and $b_j(\boldsymbol{O}_P)$ is the probability of generating prosodic features $\boldsymbol{O}_P$. λ_S and λ_P are weighting factors for the segmental stream and the prosodic stream, respectively. They are constrained by $\lambda_S + \lambda_P = 1$.

3.3.3. Building SP-HMMs

Syllable HMMs for segmental and prosodic features are separately made and combined to build SP-HMMs using a tied-mixture technique as follows:

(1) HMMs are trained by segmental features only. They are called "S-HMMs (Segmental HMMs)". They can be denoted by either "LC-SYL,*" or "SYL+RC,*". Here, "*" (wild card) means that HMMs are built without considering the F_0 transitions, "U", "D" and "F". The total number of S-HMM states is the same as SP-HMM states. Twenty S-HMMs including "sil", "sp" are trained.

(2) Training utterances are segmented into syllables by the forced-alignment technique using the S-HMMs, and one of the F_0 transition labels, "U", "D" or "F", is manually given to each segment according to the actual F_0 pattern.

(3) HMMs, each having a single state, are trained by prosodic features within these segments, according to the F_0 transition label. They are called "P-HMMs (Prosodic HMMs)", and eight separate models, "*-*,U", "*+*,U", "*-*,D", "*+*,D", "*-*,F", "*+*,F", "sil" and "sp", are made.

(4) The S-HMMs and P-HMMs are combined to SP-HMMs. Gaussian mixtures in the segmental stream of SP-HMMs are tied with corresponding S-HMM mixtures, while the mixtures in the prosodic stream are tied with corresponding P-HMM mixtures. Figure 3 shows the integration process. In this example, the mixtures of SP-HMM "i+chi,U" are tied with S-HMM "i+chi,*" and P-HMM "*+*,U".

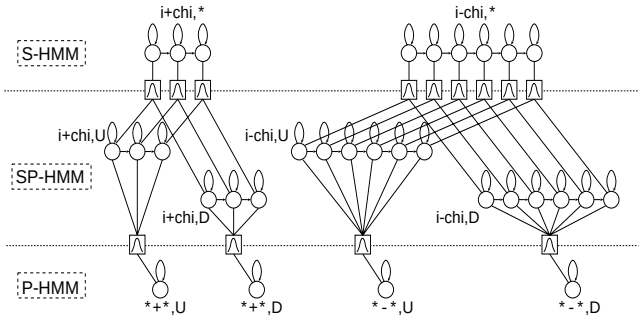


Fig. 3. Building SP-HMMs using a tied-mixture technique. S-HMMs and P-HMMs are trained using segmental features and prosodic features, respectively.

4. EXPERIMENTS

4.1. Database

A speech database was collected from 11 male speakers in a clean/quiet condition. The database consists of utterances of 2-8 connected digits with an average of 5 digits. Each speaker uttered the digit strings, separating each string with a silence period. 210 connected digits and approximately 229 silence periods were collected per speaker.

Experiments were conducted using the leave-one-out method; data from one speaker were used for testing while data from all other speakers were used for training, and this process was rotated for all speakers. Training data were clean utterances, and testing data were contaminated with either white, in-car, exhibition-hall, or elevator-hall noise at three SNR levels: 5, 10 and 20dB. Accordingly, 11 speaker-independent experiments were conducted, and a mean word accuracy was calculated as the measure of recognition performance.

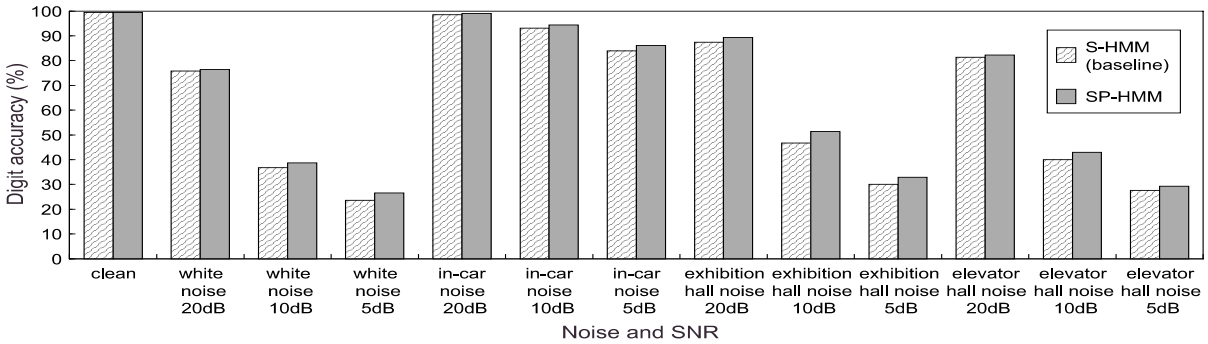


Fig. 4. Comparison of the digit recognition accuracies by SP-HMMs and S-HMMs in various noise and SNR conditions.

4.2. Dictionary and Grammar

In the recognition dictionary, each digit has three variations considering the F_0 transitions. For instance, variations of "1" are "i+chi,U i-chi,U sp", "i+chi,D i-chi,D sp", and "i+chi,F i-chi,F sp". This means that the F_0 transition pattern does not change within each digit. Recognition grammar was described so that all digits can be connected without any restriction.

4.3. Experimental Results

Training and testing were performed using HTK[5]. In our preliminary experiments, the best S-HMM recognition performance ("baseline") was obtained when the number of mixtures in each S-HMM was four. Therefore, we conducted experiments for selecting the best number of mixtures in the prosodic stream (P-HMMs) in SP-HMMs tied to the four mixture S-HMMs. The best performance of SP-HMMs was obtained when four mixture P-HMMs were used.

Figure 4 shows the digit accuracy using SP-HMMs tied to four mixture S-HMMs and four mixture P-HMMs in various noise and SNR conditions. The segmental and prosodic stream weights were optimized. Digit accuracies were improved in all kinds of noise conditions. The best improvement using SP-HMMs was observed when exhibition-hall noise was added at 10dB SNR; digit accuracy was improved by 4.7% from 46.7 % to 51.4 % by the prosodic information.

Figure 5 shows the digit recognition accuracy as a function of the prosodic stream weight λ_P when the exhibition-hall noise is added at 10dB SNR. Four mixture S-HMMs and P-HMMs were used. The best result was obtained when $\lambda_P = 0.6$. The improvement using the SP-HMMs is observed over a wide range of $0.0 < \lambda_P \leq 0.8$.

The improvement was observed for every speaker, which means that the proposed method is useful for speaker-independent recognition.

As a supplementary experiment, we compared SP-HMMs with S-HMMs in digit boundary detection capability under noisy environments. Noise-added utterances and clean utterances were segmented by these models using the forced-alignment technique. The boundary detection errors (ms) were computed by comparing the detected boundary locations in noise-added utterances with that in clean utterances. The mean digit boundary detection error rate was reduced by 17.6 % for 10dB SNR utterances and 38.2 % for 5dB SNR utterances using the SP-HMMs. We attribute this better recognition performance to the more precise boundary detection method.

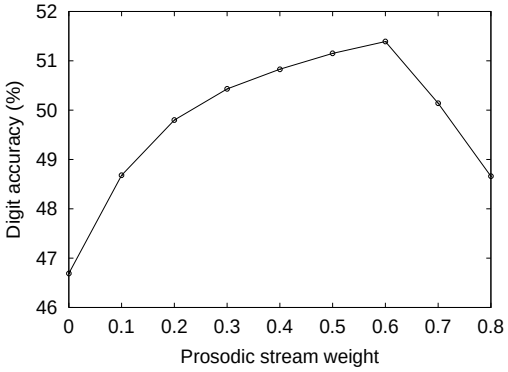


Fig. 5. Digit recognition accuracy as a function of prosodic stream weight (λ_P) when exhibition hall noise is added to speech at 10dB SNR.

5. CONCLUSIONS

This paper proposed an F_0 extraction method using Hough transform and a recognition method using syllable HMMs combining segmental and prosodic information. It was confirmed that both methods are robust in various noise conditions. Our future works include: 1) optimizing the combination of the two prosodic features, 2) investigation of the SP-HMM topology, 3) study on combination with the adaptation method such as the MLLR technique, and 4) evaluation by more general recognition tasks.

6. REFERENCES

[1] Y. Sagisaka, et al., eds., *Computing PROSODY*, part IV, Springer–Verlag, New York, 1997.

[2] T. Shinozaki, et al., "Analysis on individual differences in automatic transcription of spontaneous presentations," *Proc. ICASSP2002*, Orlando, Florida, 2002. (to appear)

[3] A. Stolcke, et al., "Modeling the prosody of hidden events for improved word recognition," *Proc. Eurospeech'99*, Budapest, vol.1, pp.311–314, 1999.

[4] P.V.C. Hough, "Method and means for recognizing complex patterns," U.S. Patent #3069654, 1962.

[5] S. Young, et al., *The HTK Book, Version 2.2*, Entropic Ltd., 1999.