論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Noise robust speech recognition using prosodic information
著者(和文)	岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, Takahiro Seki, Sadaoki Furui
出典(和文)	, Vol. , No. , pp.
Citation(English)	DSP2003, Vol. , No. , pp.
発行日 / Pub. date	2003, 4

NOISE ROBUST SPEECH RECOGNITION USING PROSODIC INFORMATION

K. Iwano, T. Seki, and S. Furui

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

ABSTRACT

This paper proposes a noise robust speech recognition method for Japanese utterances using prosodic information. In Japanese, the fundamental frequency (F_0) contour conveys phrase intonation and word accent information. Consequently, it also conveys information about prosodic phrase and word boundaries. This paper first proposes a noise robust F_0 extraction method using the Hough transform, which achieves high extraction accuracy 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. We use two prosodic features combined with ordinary cepstral parameters: a derivative of the time function of $\log F_0$ ($\Delta \log F_0$) and a maximum accumulated voting value of the Hough transform representing a measure of F_0 continuity. Speaker-independent experiments were conducted using connected digits uttered by 11 male speakers in various kinds of noise and SNR conditions. It was confirmed that both prosodic features improve the recognition accuracy in all noise conditions, and the effects are additive. When using both prosodic features, the best absolute improvement of digit accuracy is about 4.5%. This improvement was achieved by improving the digit boundary detection by using the robust prosodic information.

1. INTRODUCTION

How to increase robustness is one of the most important issues in building speech recognition systems in mobile and vehicular environments.

It has been found that human beings use prosodic information, especially the temporal pattern of fundamental frequency (F_0), to increase the robustness in recognizing speech in noisy environments. However, with the present technology, it is not easy to automatically extract correct F_0 values, especially in noisy environments. Various techniques have been proposed to smooth out incorrect values from a time series of extracted F_0 values, but these methods are not always successful. This paper proposes a novel robust method, in which the Hough transform is applied to a windowed time series of cepstral vectors extracted from speech, instead of directly extracting F_0 independently for each frame of speech. Due to its capability of extracting straight-line components from an image, the Hough transform can extract a reliable F_0 value for each window. By shifting the window at every frame, a smooth time function of F_0 can be obtained.

We also propose a speech recognition method using prosodic features extracted by the Hough transform, consisting of a derivative of the time function of $\log F_0$ ($\Delta \log F_0$) or/and a measure of periodicity. These features are combined with ordinary cepstral parameters and modeled by multi-stream HMMs, which are trained using clean speech. Since F_0 contours represent phrase intonation and word accent in Japanese utterances, prosodic features are useful to detect prosodic phrases and word boundaries. Therefore, the proposed method using robust prosodic information is able to precisely detect word boundaries and improve recognition performance under noisy environments.

The paper is organized as follows. In Section 2, a robust F_0 extraction method using the 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[1].

The Hough transform method to extract a significant line from an image on the 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 for all the 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

Cepstral peaks extracted independently for each short period of speech have been widely used to extract F_0 values. This method often causes errors, including half pitch, double pitch and drop outs, for noisy speech. Since F_0 contours have temporal continuity in voiced periods, the Hough transform, taking advantage of its continuity, applied to time-cepstrum images is expected to have robustness in extracting pitch in the noisy environment.

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, time continuity for 90ms is considered in this method.

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 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 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

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 the 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 using 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.

Two prosodic features are computed: one is the $\Delta \log F_0$ value which represents the F_0 transition, and the other is the maximum

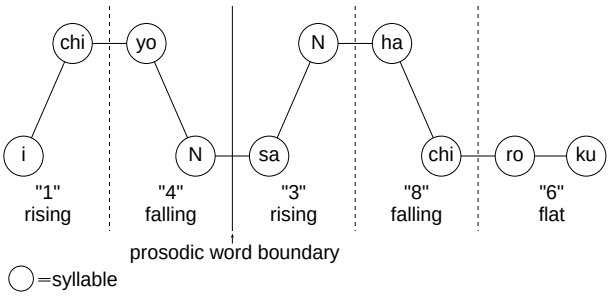
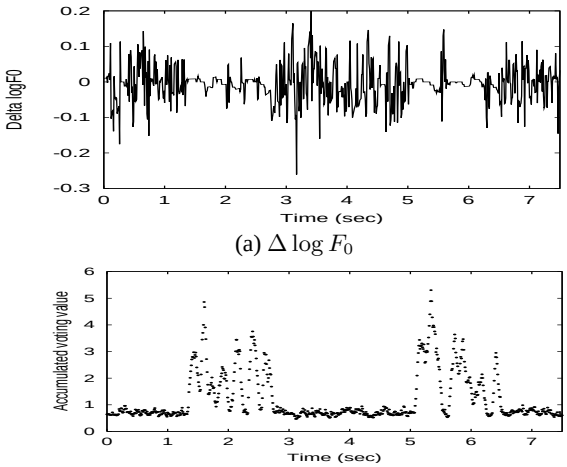


Fig. 1. An example of F_0 contour of Japanese connected digit speech.



(a) $\Delta \log F_0$

(b) Maximum 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.

accumulated voting value obtained in the Hough transform which indicates the degree of temporal continuity in the F_0.

An example of the time function of $\Delta \log F_0$ and maximum 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.

In this paper, two kinds of prosodic features and their combination, **P-D**, **P-V**, and **P-DV**, are investigated:

P-D: $\Delta \log F_0$

P-V: maximum accumulated voting value

P-DV: $\Delta \log F_0$ + maximum accumulated voting value

These three kinds of prosodic features are combined with segmental features for each frame. Therefore, three kinds of segmental-prosodic feature vectors are built and evaluated.

3.3. Multi-stream Syllable HMMs

3.3.1. Basic Structure of Syllable HMMs

Since syllable transition and the change of F_0 characteristics such as "rising", "falling" and "flat" are highly related, the segmental

and prosodic features are integrated using syllabic unit HMMs. Our preliminarily experiments showed that the syllable unit HMMs have approximately the same digit recognition accuracy for a connected digit task as tied-state triphone HMMs.

The integrated syllable HMM denoted by "SP-HMM (Segmental-Prosodic HMM)" models both phonetic context and F_0 transition. Each Japanese digit uttered continuously with other digits can be modeled by a concatenation of two context-dependent syllables. Even "2" (/ni/) and "5" (/go/) can be modeled by two syllables since their final vowel is often lengthened as /ni:/ and /go:/. The context of each syllable is considered only within each digit in our experiment. Therefore, each SP-HMM is denoted by either a left-context dependent syllable "LC-SYL,PM" or a right-context dependent syllable "SYL+RC,PM", where "PM" indicates a 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. "sil" and "sp" models are used for representing 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 the recognition stage, the probability $b_j(\boldsymbol{O}_{SP})$ of generating segmental-prosodic observation $\boldsymbol{O}_{SP}$ at state j is calculated by:

$$b_j(\boldsymbol{O}_{SP}) = b_j(\boldsymbol{O}_S)^{\lambda_S} \cdot b_j(\boldsymbol{O}_P)^{\lambda_P} \tag{3}$$

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 and prosodic streams, 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) "S-HMMs (Segmental HMMs)" are trained by using only segmental features. They are denoted by either "LC-SYL,*" or "SYL+RC,*". Here, "*" (wild card) means that HMMs are built without considering the F_0 transitions, "U", "D" or "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 then, one of the F_0 transition labels, "U", "D" or "F", is manually given to each segment according to its actual F_0 pattern.

(3) "P-HMMs (Prosodic HMMs)", having a single state, are trained by prosodic features within these segments, according to the F_0 transition label. 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 for the segmental feature stream of SP-HMMs are tied with corresponding S-HMM mixtures, while the mixtures for the prosodic feature stream are tied with corresponding P-HMM mixtures. Figure 3 shows the integration process. In this example, mixtures of SP-HMM "i+chi,U" are tied with those of S-HMM "i+chi,*" and P-HMM "*+*,U".

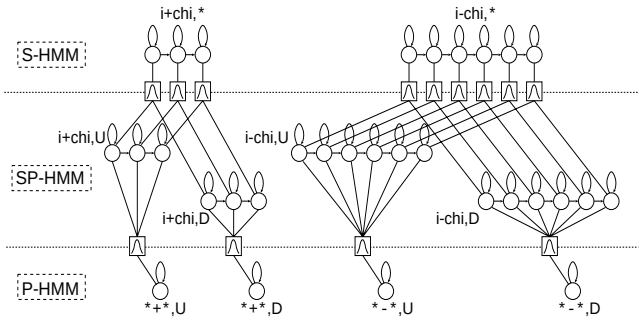


Fig. 3. Building SP-HMMs using a tied-mixture technique. S-HMMs and P-HMMs are trained using segmental and prosodic features, respectively.

4. EXPERIMENTS

4.1. Database

A speech database was collected from 11 male speakers in a clean/quiet condition. The database comprised 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 each speaker. Accordingly, 11 speaker-independent experiments were conducted, and a mean word accuracy was calculated as the measure of recognition performance. All the HMMs were trained using only 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.

4.2. Dictionary and Grammar

In the recognition dictionary, each digit had three variations considering the F_0 transitions. For instance, variations of "1" comprised "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 was not allowed to change within each digit. Recognition grammar was described so that all digits could be connected without any restriction.

4.3. Experimental Results

Training and testing were performed using the HTK[2]. In our preliminary experiments, the best S-HMM recognition performance ("baseline") was obtained when the number of mixtures in each S-HMM was four. Experiments for selecting the optimum number of mixtures for the prosodic stream (P-HMMs) in SP-HMMs tied

Table 1. Digit recognition accuracies by SP-HMMs and S-HMMs in various SNR conditions.

SNR	**S-HMM** (baseline)	**SP-HMM -D**	**SP-HMM -V**	**SP-HMM -DV**
clean	99.3	99.6	99.4	99.4
20 dB	84.9	86.0	85.7	86.1
10 dB	53.1	54.6	55.1	55.7
5 dB	40.1	41.4	42.2	42.7

with four mixture S-HMMs were conducted, and the best performance using SP-HMMs was obtained when four mixture P-HMMs were used. Therefore, in the experiments hereafter, SP-HMMs were tied with four mixture S-HMMs and four mixture P-HMMs.

Table 1 shows the digit accuracy using SP-HMMs in various SNR conditions. "**SP-HMM-X**" indicates the SP-HMMs using the prosodic feature "**P-X**". Accuracies for four kinds of noises are averaged at 20, 10, and 5dB SNR, respectively. The segmental and prosodic stream weights and insertion penalties were optimized for each noise condition. Digit accuracies were improved in all kinds of noise and prosodic feature conditions. It can be seen that **SP-HMM-DV** showed the best performance, which means that the effects of the $\Delta \log F_0$ and the maximum accumulated voting value are additive. The best improvement of 4.5% from 45.3 % to 49.8 % is observed in the condition when exhibition-hall noise was added at 10dB SNR and the prosodic feature **P-DV** was used.

Detailed results showed that the improvement was achieved for every speaker, which means that the proposed method is effective in speaker-independent speech recognition.

Figure 4 shows the improvement of digit recognition accuracy as a function of the prosodic stream weight λ_P at each SNR. Results for four kinds of noises are averaged at 20, 10, and 5dB SNR, respectively. In this experiment, the prosodic feature **P-DV** was used, and insertion penalties were optimized. The improvement using the SP-HMMs was observed over a wide range: $0.0 < \lambda_P \leq 0.7$ in all the noise conditions. Best results were obtained when λ_P was set aroung 0.6, irrespective of the SNR level.

Figure 5 shows the optimum insertion penalty as a function of the prosodic stream weight λ_P in white noise condition, when the prosodic feature **P-DV** was used. In noisy conditions, if the prosodic stream weight is low, we need to set the insertion penalty high to compensate for the low reliability of segmental features. Since prosodic features are effective for digit boundary detection, the higher the prosodic stream weight becomes, the lower the optimum insertion penalty becomes. Similar results were obtained for other noise conditions. The control range of the optimum insertion penalties in the best prosodic stream weight condition ($\lambda_P = 0.6$) is approximately a half of the range for the condition without using the prosodic information. This means that the prosodic features are effective for robust adjustment of the insertion penalty.

As a supplementary experiment, we compared the boundary detection capability of SP-HMMs and S-HMMs in digit recognition under noisy environments. Noise-added utterances and clean utterances were segmented by both of these models using the forced-alignment technique. The boundary detection errors (ms) were measured 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 23.2% for 10dB SNR utterances and 52.2% for 5dB SNR utterances using the **SP-HMM-DV**. These results indicate the effectiveness of prosodic information in digit boundary detection.

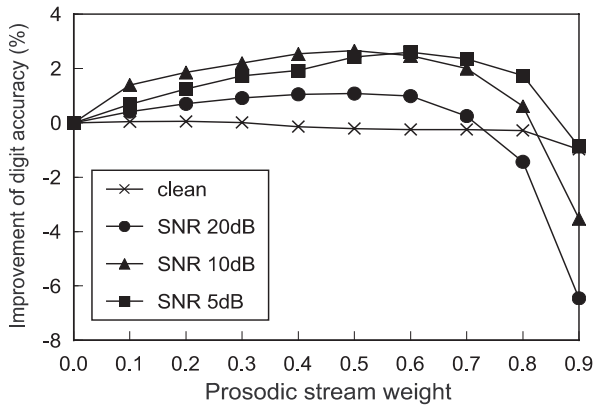


Fig. 4. Improvement of digit accuracy as a function of prosodic stream weight (λ_P) in each SNR condition.

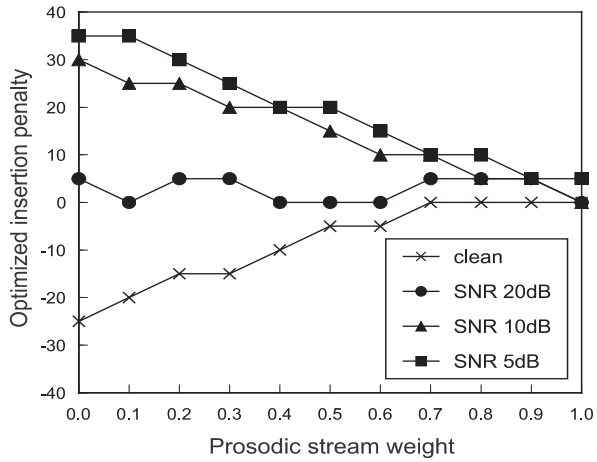


Fig. 5. Optimized insertion penalty as a function of prosodic stream weight (λ_P) in white noise condition.

5. CONCLUSIONS

This paper has proposed an F_0 extraction method using the Hough transform and a new speech recognition method using syllable HMMs utilizing both segmental and prosodic information. Both methods were confirmed to be robust in various noise conditions. The prosodic information is effective in digit boundary detection and consequently improves connected digit recognition performance under noise. Future works include combination of our method with model adaptation or feature normalization techniques for noise effects and evaluation using more general recognition tasks.

6. REFERENCES

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

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