Speech Estimation Biased by Phonemic Expectation in the Presence of Non-stationary and Unpredictable Noise

Ikuyo Masuda-Katsuse and Yoshimori Sugano

Kyushu Institute of Design / Institute of Systems & Information Technologies
2-1-22, Momochihama, Sawara-ku, Fukuoka City, 814-0001, Japan.
{ikuyo, sugano}@isit.or.jp

Abstract

In this paper, we propose a new method for speech recognition in the presence of non-stationary and unpredictable noise by extending PreFEst [4]. The method does not need to know noise characteristics in advance and does not even estimate them in its process. A small set of pre-evaluations demonstrates the feasibility of the method by demonstrating good performance with a signal-to-noise ratio of 10 dB.

1. Introduction

In establishing a framework for robust speech recognition in the usual noisy environment, at least the following two requirements should be considered: First, that prior information on background noise or its modeling is not required, because background noise is usually non-stationary and unpredictable; and second, that it is well connected with conventional automatic speech recognition (ASR) techniques. The missing data approach proposed by Cooke et al. satisfies both of those requirements [1]. In the Cooke et al. approach, the auditory spectrogram of the speech source is identified by using computational primitive auditory scene analysis. Next, the unreliable parts of the spectrogram are identified. In the marginalization approach, speech recognition is performed based solely on the reliable data. In the data imputation approach, unreliable parts are estimated using hidden Markov Model (HMM) state distributions. This approach might be further considered as an attempt to combine the bottom-up and top-down processes of speech recognition.

With human speech perception, top-down and bottom-up processes are used together to perceive speech. However, the top-down process is so active in generating ”expectation” of the existence of a particular speech that it biases the processing mechanism so as to produce an answer consistent with the expectation, whereas a passive process compensates for the incomplete part of the bottom-up processing result using linguistic knowledge [2].

In this paper, we propose a framework for speech recognition that permits the active intervention of the top-down process in speech estimation by extending the PreFEst (Predominant-F0 Estimation Method) [3] [4]. The conventional ASR possesses standard spectral envelope data. By making use of these data as top-down information in our process, our process can be incorporated into the conventional ASR.

2. Outlines

2.1. Extension of PreFEst[3][4]

The PreFEst represents spectral components of the observed signal as a probability density function (PDF). This PDF is generated from a weighted-mixture model of tone models for all possible fundamental frequencies. The PDF of the m-th tone model whose fundamental frequency is F is described as follows:

$$p(x|F,m,\mu^{(t)}(F,m)) = \sum_{h=1}^{H} p(x,h|F,m,\mu^{(t)}(F,m)) \quad (1)$$

where

$$p(x,h|F,m,\mu^{(t)}(F,m)) = c^{(t)}(h|F,m)G(x|F,h), \quad (2)$$

H denotes the number of harmonics taken into consideration, and G denotes the Gaussian distribution that has its maximum at $h \cdot F$. The observed PDF $p_{\psi}^{(t)}(x)$ is generated from a weighted-mixture model $p(x|\theta^{(t)})$ of $p(x|F,m,\mu^{(t)}(F,m))$,

$$p(x|\theta^{(t)}) = \int_{Fl}^{Fh} \sum_{m=1}^{M} w^{(t)}(F,m)p(x|F,m,\mu^{(t)}(F,m))dF \quad (3)$$

where Fl and Fh denote the lower and upper limits of the possible fundamental frequencies, respectively, and $w^{(t)}(F,m)$ denotes the weighting value of a tone model.

We now have phonemic knowledge (that is, standard spectral envelope information) as prior knowledge, and want to estimate the model parameters based on prior distribution generated by using the prior knowledge.

To estimate parameter $\theta^{(t)}$ of model $p(x|\theta^{(t)})$ of the observed PDF based on prior distribution $p_0(\theta^{(t)})$, each iteration in the EM algorithm updates the old estimate $\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$ to obtain the new improved estimate $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$ for estimation of the maximum a posterior probability of $\theta^{(t)}$ based on the prior distribution (for details see Goto[4]).

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{ML}^{(t)}(F,m)} + \beta_{\omega}^{(t)}\omega_0^{(t)}(F,m)}{1+\beta_{\omega}^{(t)}} \quad (4)$$

and

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{\omega_{ML}^{(t)}(F,m)c_{ML}^{(t)}(h|F,m)} + \beta_{\mu}^{(t)}(F,m)c_0^{(t)}(h|F,m)}{\overline{\omega_{ML}^{(t)}(F,m)} + \beta_{\mu}^{(t)}(F,m)} \quad (5)$$

where $\overline{w_{ML}^{(t)}(F,m)}$ and $\overline{c_{ML}^{(t)}(h|F,m)}$ are the maximum likelihood estimates when noninformative prior distribution is given, and $\omega_0^{(t)}(F,m)$ and $c_0^{(t)}(h|F,m)$ are the most probable parameters given in advance.

$c_0^{(t)}(h|F,m)$ is generated using standard speech envelopes $C_0^{(t)}(x|m)$ possessed by ASR as follows:

$$c_0^{(t)}(h|F,m) = \frac{C_0^{(t)}(x|m)\delta(h\cdot F)}{\sum_{h=1}^{H} C_0^{(t)}(x|m)\delta(h\cdot F)}. \quad (6)$$

$\beta_\omega^{(t)}$ determines how much importance is attached to $\omega_0^{(t)}(F,m)$, and $\beta_\mu^{(t)}(F,m)$ determines how much importance is attached to $c_0^{(t)}(h|F,m)$.

”Expected” phoneme can be preferentially estimated by allocating more weighting values to the standard spectral envelope and pitch frequency of the speech whose existence is expected from the linguistic context and prosody context, and so on.

Because we want to know not the dominant tone model but the dominant $C_0^{(t)}(x|m)$, we define the dominance of $C_0^{(t)}(x|m)$ in the observed signal as follows:

$$W_d^{(t)}(m) = \int_{Fl}^{Fh} \omega^{(t)}(F,m)dF. \quad (7)$$

Here, we note that although we make use of the harmonic structure of speech for estimating, we do not have to estimate its fundamental frequency in advance.

2.2. Unvoiced Consonant Model

The tone model presupposes that the target speech always has a harmonic structure. However, because unvoiced speech sounds do not have harmonic structures, whether the tone model is suitable or not as an unvoiced consonant model should be tested. We compared three types of models should be tested. One is a tone model similar to that for voiced speech. Another is an envelope model that has a smooth and continuous spectral envelope. The other is a tone model with constant component amplitudes. Each weighted value of the target speech estimated by using each unvoiced consonant model is compared. As a result, the greatest weighted value was observed in the tone model similar to that for voiced speech. The tone models of both voiced speech and unvoiced speech are therefore made with equation (6).

2.3. PDF of Tone Model and Observed Signal

Standard speech envelopes are generated from a 16-element MFCC (Mel-Frequency Cepstrum Coefficient). The envelopes are represented on the mel-frequency axis.

The variance of Gaussian distribution used as an F0-dependent weighting function that emphasizes regions near the harmonics is constant on the mel-frequency axis. Therefore, the allowable frequency-error range of a harmonic component increases as the frequency increases. The PDF that represents the spectral components of the observed signal is also mapped on the mel-frequency axis.

The sampling frequency is 16 kHz, and the number of iteration in the EM algorithm is five.

3. Pre-Evaluation for Application to ASR

3.1. Noisy Speech for Evaluation

A noisy environment was simulated in an anechoic room using factory noise recorded in the JEIDA noise database [6]. The noise was non-stationary. One male speaker uttered eight names of prefectures in Japan five times, at a level somewhat louder than usual in the noisy environment. The speech sounds through a headset microphone (B&K Type4035) and those through a throat microphone (Audio-Technica AT890) were recorded into the L and R channels of a digital audio tape recorder, respectively.

The noise level at the location of the headset microphone was adjusted to the same level as the noise was recorded. As a result, the noise level was about 70 dB(C). Because the sound pressure level of the speech was about 78 dB(C), the signal to noise ratio was about 8 dB. In addition to the noisy speech, clean speech sounds were recorded five times. One set of clean speech sounds was used as the standard speech.

3.2. Determination of Speech Period

It is difficult, and it often decisively influences the performance of speech recognition, to determine the speech period in the presence of high-level noise. In an HMM-based ASR, speech period extraction is avoided by connecting the noise HMM before and after the speech HMM. In this case, the noise model must be presupposed beforehand, which is not realistic when the background noise is non-stationary and unpredictable. In the case of the dynamic programming (DP)-matching-based ASR, an unconstrained endpoint DP-matching algorithm has been proposed. However, calculation cost remains a problem.

The speech period is detected by sensing the glottal sound directly using a contact microphone, such as a throat microphone. When we use the speech-input device under high-level noise, it is practical to use a portable or wearable near-microphone such as a headset microphone; therefore, the use of the contact microphone does not seem to annoy the speaker. We can determine the voiced period by watching the time change of the zero-crossing count of the short-term glottal sound, because the count provides a cue for whether the vocal cord vibrates or not. Although the unvoiced parts at the start or end of the utterance cannot be included in the speech period detected by using the cue, this does not pose a serious problem because the signal-to-noise ratio at the unvoiced part is often low in the presence of high-level noise.

All of the clean or noisy speech sounds for the evaluations described in this section were automatically extracted from recorded tapes by using this method.

3.3. DP-matching based on Dominance

The tone models that compose a weighted-mixture model at time t are prepared by using several pieces of spectral envelopes included in the adjustment time window from each standard speech. The cumulative summation of the dominance, which is called ”the word dominance”, is obtained using the DP-matching algorithm [5] based on the dominance of $C_0^{(t)}(x|m)$ at each time.

3.4. Word Dominance of Candidate Word

Figure 1 shows the mean word dominance of candidate words for noisy speech recorded in the quasi-noisy room. The horizontal axes show the spoken words and the vertical axes show the

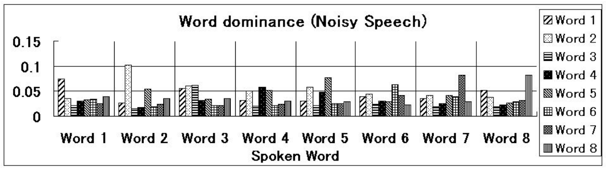

Figure 1: *Word dominance for noisy speech recorded in noisy environment*

word dominance of the candidate words. The results demonstrate that the dominance of the correct word is greater than those of the other words.

3.5. Relation between Word Dominance and Signal to Noise Ratio

We created some noisy speech artificially by adding a part of the factory noise to clean speech at several signal to noise ratios (SNR) to evaluate the difference in performance according to the SNR. Figure 2, 3, 4 and 5 show the results obtained for clean speech and several kinds of noisy speech. The results show that the word dominance of the correct word and the margins (namely, the differences from those of the other words) become smaller as the SNR decreases.

4. Application to Discrete-HMM-based ASR (DHMM-ASR)

4.1. Incorporating the Proposed Method into DHHM-ASR

Figure 6 shows an outline of a general Discrete-HMM-based ASR(DHHM-ASR) for word recognition. A short-term segment of speech is frequency-analyzed and its spectrum is represented by MFCC that are vector-quantized using a codebook. The recognition word is determined using the quantized vector and the word-level HMMs with the maximum likelihood method.

Combining the proposed method with a DHMM-ASR is achieved by replacing the part enclosed by a broken line in Fig. 6 with the whole presented in Fig. 7. The standard spectral envelopes used as prior knowledge in the proposed method are mapped into centroids on the codebook used in the DHHM-ASR. By regarding the centroid whose dominance is the maximum as an observed code, decoding along the conventional HMM-based ASR algorithm will be possible.

In this way, the proposed method will be used as a part of an HMM-based ASR without great modification or re-learning of the model.

5. Discussion

We proposed a new framework that is an extension of the PreFEst to achieve speech recognition in the presence of nonstationary and unpredictable noise. A small set of evaluations shows the feasibility of the proposed method.

The proposed method has parameters to estimate ”expected” phonemes or fundamental frequencies preferentially, although the effective use of the parameters has not been achieved in this report. Several psychological studies[7][8][9][10] have shown the possibility of improved performance of the ASR in the presence of noise by effectively using information about linguistic context or prosody context, and so on.

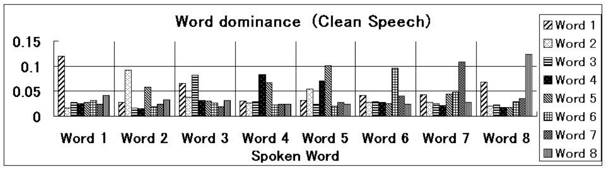

Figure 2: *Word dominance for clean speech*

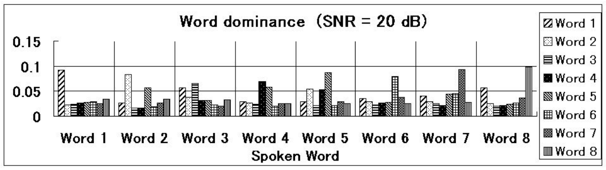

Figure 3: *Word dominance for noisy speech (SNR = 20 dB)*

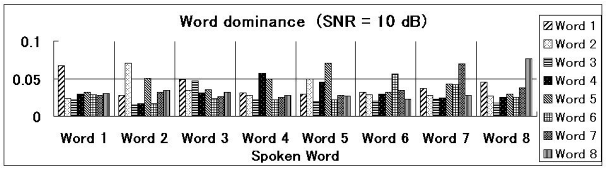

Figure 4: *Word dominance for noisy speech (SNR = 10 dB)*

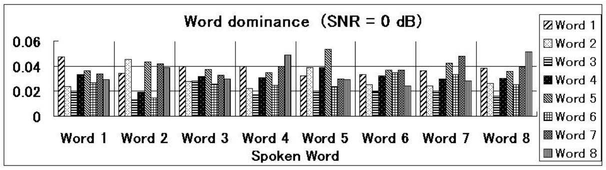

Figure 5: *Word dominance for noisy speech (SNR = 0 dB)*

As we mentioned, it is important that a presupposition about background noise is not indispensable. However, if noise characteristics are available, they may be used for estimation of the target speech. In the proposed method, the information about background noise can be used by adding to a weighted-mixture model as a noise model.

Finally, we note the relationship between the proposed method and a speech perception model. de Cheveigné and Kawahara proposed a missing-data model of vowel identification. They treated vowel identification as a process of pattern recognition, and data in matching were restricted to regions near harmonics. Our proposed method corresponds to an extended example of the frequency-domain version of their model because, in our proposed method, whether or not representations obtained by emphasizing the spectral envelopes of speech at multiples of the F0 seems to be contained in the observed signal is estimated.

Acknowledgement
This work is supported by WaveCom Inc.

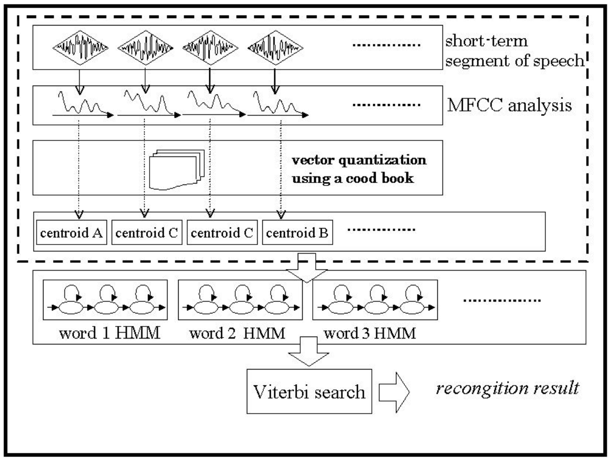

Figure 6: *Outline of a general discrete HMM for word recognition.*

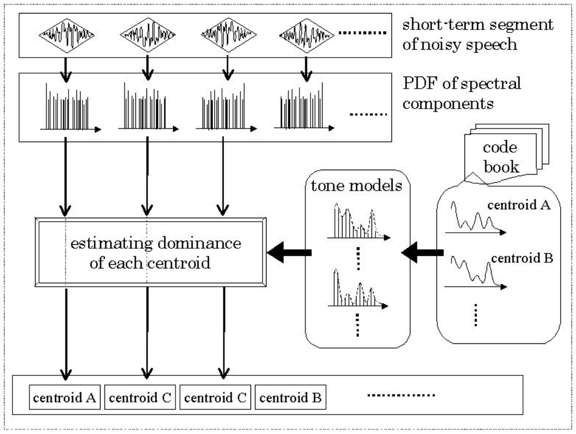
short-term segment of noisy speech
PDF of spectral components
code book
tone models
centroid A
centroid B
estimating dominance of each centroid
centroid A
centroid C
centroid C
centroid B

Figure 7: *One method for introducing the proposed method into DHMM-ASR.*

6. References

[1] Cooke, M. et al., ”Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Communication 34(3), 2001.

[2] Tohkura, Y., ”Section 4: Auditory mechanism and perception model,” in *Speech, Audition and Neural Network Model*, ed. by Amari, S., Ohmsha, 1990. (in Japanese)

[3] Goto, M., ”A robust predominant-f0 estimation method for real-time detection of melody and bass lines in CD recordings,” Proc. of ICASSP2000:II-757–760, 2000.

[4] Goto, M., ”A predominant-F0 estimation for CD recordings: MAP estimation using EM algorithm for adaptive tone models”, Proc. of ICASSP2001, 2001.

[5] Sakoe, H. and Chiba, S., ”Dynamic programming algorithm optimization for spoken word recognition,” IEEE ASSP-26(1): 43–49, 1978.

[6] Itahashi, S., ”A noise database and Japanese common speech data corpus,” J. Acoust. Soc. Jpn., Vol. 47(12): 951–953, 1991. (in Japanese)

[7] Broadbent, D. E., ”Word-frequency effect and response bias,” Psychol. Review, 74: 1–15, 1967.

[8] Miller, G. A., et al., ”The intelligibility of speech as a function of the context of the test materials,” J. Experimental. Psychol., 41: 329–335, 1951.

[9] Samuel, A.G., ”Lexical uniqueness effects on phonemic restoration,” J. Mem. Lang. 26: 36–56, 1987.

[10] Samuel, A.G., ”A further examination of attentional effects in the phonemic restoration illusion,” Q. J. Exp. Psychol. 43A: 679–699, 1991.

[11] de Cheveigné, A. and Kawahara, H., ”Missing-data model of vowel identification,” J. Acoust. Soc. Am., Vol. 105(6): 3497–3508, 1999.