International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:6, No:10, 2012

Puff Noise Detection and Cancellation for Robust Speech Recognition

Sangjun Park, Jungpyo Hong, Byung-Ok Kang, Yun-keun Lee, and Minsoo Hahn

***Abstract*—**In this paper, an algorithm for detecting and attenuating puff noises frequently generated under the mobile environment is proposed. As a baseline system, puff detection system is designed based on Gaussian Mixture Model (GMM), and 39th Mel Frequency Cepstral Coefficient (MFCC) is extracted as feature parameters. To improve the detection performance, effective acoustic features for puff detection are proposed. In addition, detected puff intervals are attenuated by high-pass filtering. The speech recognition rate was measured for evaluation and confusion matrix and ROC curve are used to confirm the validity of the proposed system.

***Keywords*—**Gaussian mixture model, puff detection and cancellation, speech enhancement.

I. Introduction

RECENTLY, speech recognition in electrical devices engages public attention with wide spread of smart phones and tablet PCs. Speech recognition in the condition that a microphone is adjacent to user's mouth can have a problem with intermittent noises. A puff noise, one of the intermittent noises, is a major problem that has harmful effects on the speech recognition system in the condition.

The puff noise is caused by a mass of unintended air inflow coming from user's mouth or nose. It is mainly occurred by user's utterance habit, and the noise is recognized as an additional syllable because it is closely located to speech signals. Aspirated noise and wind noise are similar with puff noise, but the characteristics are rather different. Normally, aspirated noises indicate a turbulent noise produced by a sufficiently narrow constriction at the glottis in speaking [2]. On the other hand, duration of wind noises is relatively continuous compared with the others. The differences are summarized in Table I.

Many researchers have been studied to detect and reduce this kind of noises. In [3]-[6], systems for wind noise reduction are proposed using single microphone. Speech enhancement methods based on non-negative matrix (NMF) are introduced in [3], [4]. They are effective for wind noise reduction, but excessive computation time is required. In addition, spectral subtraction and template matching techniques detection based on formant invariant characteristics that are used to obtain noise power spectrum [5]. Wind noise independent to variation of order of linear prediction coefficients (LPC) is utilized in [6]. Besides, Multi-channel techniques are introduced in [7], [8]. The approaches based on noise power estimation are not proper to puff noise detection and cancellation because the noise statistics varies rapidly and noise occurrence is sporadic.

TABLE I
Classification of Analogous Terms

Classification Category / Noise Type		Location	
		Non-Overlap	Overlap
Duration	Occasional	Puff noise	Aspirated noise
	Continuous	Wind noise	

In this paper, probabilitistic puff detection using Gaussian mixture model (GMM) is proposed. It is two-class classification between speech and puff noise. Acoustic features effective for puff noise detection are proposed to replace 39th mel-frequency cepstral coefficients (MFCC) which are features of a baseline system. The proposed features have lower dimension with high performance in detecting puff noises. After the detection, the intervals regarded as puff interval are attenuated by high-pass filter.

The paper is organized as follows. Puff detection system of baseline and proposed one is explained in section 2. Then, puff cancellation is briefly mentioned in section 3. In section 4, experiments and results will be shown. Finally, conclusion follows.

II. Puff noise detection

To modeling the puff noises, general approach based on Gaussian mixture model (GMM) is used. Puff noise detection is same to the problem that two class classification between clean speech and puff noises. The proposed puff detection system is shown in Fig. 1. It is composed of two processes: offline and online. The offline process is designed to train GMM parameters of each class, means and variances of speech and puff noises, and the parameters are provided to GMM classifier in order to decide the present frame belongs to puff noise or clean speech. Finally, it is decided that filtering is performed or not, according to the detection results.

Sangjun Park and Jungpyo Hong are with the Department of Electrical Engineering, KAIST, Daejeon, Republic of Korea (e-mail: psj@kaist.ac.kr, e-mail: hansin@kaist.ac.kr).

Byung-Ok Kang and Yun-keun Lee are with Electronics and Telecommunications Research Institute, Republic of Korea (e-mail: bokang@etri.re.kr, yklee@etri.re.kr).

Minso Hahn is with the Department of Electricial Engineering KAIST Daejeon Republic of Korea (phone: +82-42-350-3474; fax: +82-42-350- 7619; e-mail: mshahn2@kaist.ac.kr), corresponding author.

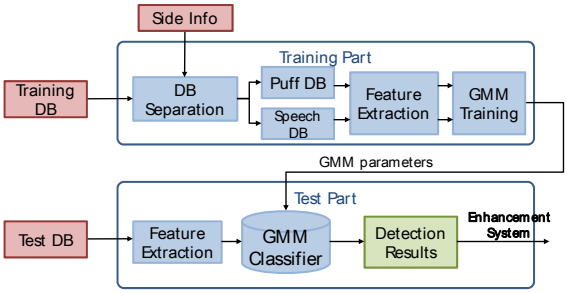

Fig. 1 Block diagram of puff detection system

A. Baseline System with MFCC

MFCC is a typical feature parameter in speech signal analysis. It is based on the fact that human auditory system is more sensitive to low frequency components. Thus, Mel scale filterbank is densely designed at the low frequency ranges. 39th MFCCs feature vectors are extracted from each frame for the baseline detection system. Table II shows the specific configuration about MFCC feature extraction.

TABLE II
CONFIGURATION OF MFCC FEATURE EXTRACTION

Configuration	
Number of Cepstral Coefficients with Energy	13
delta	13
delta-delta	13
Window	Hamming
Frame size	20 ms
Shift size	10 ms
Pre-emph	0.97
Format	HTK MFCC

B. Proposed System with Acoustic Features

The distinguishing characteristics of puff noises are that most of the noise power is concentrated on the low frequency range, that it has relatively large power but does not have harmonic structure. Based on the spectral or temporal characteristics of puff noise, acoustic features for effective puff detection with lower dimension than the baseline system are proposed.

1. Low-Band Energy Ratio

As mentioned above, the spectral energy of puff noise is concentrated on the low-frequency band. Thus, the low band energy ratio, low frequency band power divided by entire band power, is a prominent feature for distinguishing the noise from unvoiced speeches.

$$L-BER = \frac{\sum_{k=0}^{L} \log(|X(k)|)}{\sum_{k=0}^{N-1} \log(|X(k)|)} \quad (1)$$

where L, N, $X(k)$ are the FFT index of cuff-off frequency, FFT size and spectrum of the input speech signal, respectively.

2. Periodicity

Low-band energy ratio is effective to discriminate unvoiced speech from puff noises, however, voiced speeches also have high low band energy ratio. In consideration of the low periodicity in frequency-domain, periodicity is calculated. The feature is computed as the difference of first peak-valley of spectral autocorrelation function. Spectral autocorrelation function is obtained using only low-band spectrum in order to emphasize periodicity of voiced speeches.

3. Spectral Flux

Generally, spectral flux is a measure of how quickly the power spectrum of a signal is changing. In case of puff noises, spectrum changes rapidly by the flow turbulences and fluctuations in contrast to the slowly varying speech signal. The spectral flux is represented by,

$$S\_Flux(m) = \sqrt{\sum_{k=0}^{N-1} (X(m,k) - X(m-1,k))^2} \quad (2)$$

where m, k, N, $X(m,k)$ denote the frame index, spectral bin index, the number of spectral bin and magnitude of k^{th} spectral bin, respectively.

4. Formant Characteristics

Different from that of speech signals, formant structure of wind noises has the only formant independent to variation of LPC order [6]. The formant-invariant characteristic of puff noises can be parameterized by measuring cross correlation of LPC envelops which has different orders. The feature values tend to be high in the puff noise intervals.

5. Spectral Tilt (A1-A3)

With the formant-invariant characteristic, the spectral tilt (A1-A3) is also calculated. The spectral tilt is computed as the difference between the powers of the first formant and the third formant. A1 and A3 are estimated by the largest powers from 100 to 1000Hz and from 1800 to 4000Hz using robust formant tracking [1].

6. Mid-band Spectral Flatness

The spectrums of the puff noise from 1800 to 4000 Hz are relatively flat compared to other frequency band because of the absence of formant. Thus, the mid-band spectral flatness measure is expected to be high in the puff interval. Mid-band spectral flatness measure is denoted as

$$M\_SFM = \frac{\sqrt[N]{\prod_{k=0}^{N-1} X(k)}}{\frac{\sum_{k=0}^{N-1} X(k)}{N}} \quad (3)$$

7. Band Envelop Correlation Index (f1f3 sync)

Aspiration noises affect on around the third formant region from 1800 Hz to 4000 Hz. Thus, a cross correlation between band-pass filtered envelop of low frequency band and that of frequency band corrupted by aspiration noises is expected to be low [2].

8. Brightness and Bandwidth

The brightness is the frequency centroid of the spectrum in a frame, it can be defined as

$$SC(m) = \frac{\sum_{k=0}^{N-1} k\,|X(k)|^2}{\sum_{k=0}^{N-1} |X(k)|^2} \quad (4)$$

Bandwidth is the square root of the power-weighted average of the squared difference between the spectral components and frequency centroid, it is represented by,

$$BW(m) = \sqrt{\frac{\sum_{k=0}^{N-1} (k - SC(m))^2\,|X(k)|^2}{\sum_{k=0}^{N-1} |X(k)|^2}} \quad (5)$$

Brightness and bandwidth represent the frequency characteristic, and they have shown effectiveness in many audio classification systems [9].

9. Zero Crossing Rate

The zero crossing rate is a typical time domain feature and it is computed by

$$ZCR = \frac{1}{L-1}\sum_{k=1}^{L-1} \frac{1}{2}\,|\,\mathrm{sgn}(x(k)) - \mathrm{sgn}(x(k-1))| \quad (6)$$

where L, k, sgn function are frame length, time index and signum function. The puff noise has low zero crossing rate whereas that of unvoiced speech is high.

10. Spectral Roll-off Point

The spectral roll-off point is also typical frequency domain feature which means 85% point of overall spectrum power. Thus, puff noise which has strong low frequency component has low spectral roll-off point.

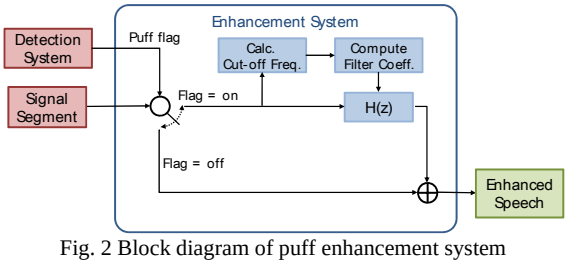

Fig. 2 Block diagram of puff enhancement system

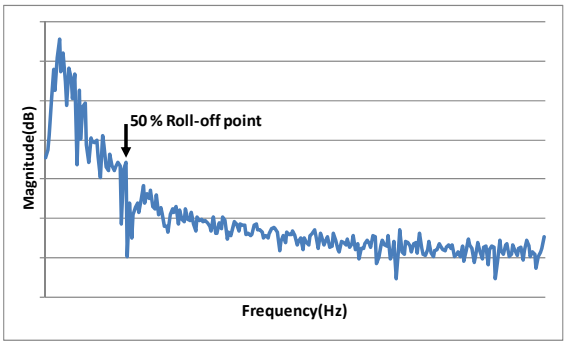

Fig. 3 Spectrum of puff and spectral roll-off point

III. Puff Noise Cancellation

Proposed puff cancellation system is an adaptive high-pass filtering. Cut-off frequency of the filter is adaptively adjusted depending on the spectral-roll off point where 50% energy is concentrated on. The system is represented in Fig. 2. Detection result in Fig. 1 is transferred to the system. When the status of puff flag is "on", high-pass filtering is performed. Fig. 3 shows an example of spectrum of puff noises.

IV. Experiments and Results

A. Configuration of Experiments

The database used for evaluation is collected by mobile speech recognition application. The length of each sample is about 2 second and puff noise is about 0.1~0.3 second in noisy samples. Each samples include 1~3 words. The features are computed with 20 ms window shifting 10 ms.

Hidden Markov Model Toolkit (HTK)-based speech recognizer was used for evaluating puff enhancement system [10]. Triphone-based HMMs is used with 1000 tied states. 39th-order MFCC was used as the feature vector.

250 clean speech samples and 600 noisy samples corrupted by puff noise are used in experiment. 200 clean speech samples and 175 noisy samples are used for training GMM. The location of puff noise is manually labeled to extract from noisy speech samples. 50 clean speech samples and 100 noisy samples are used to evaluate the puff detection system, and 325 noisy samples are used to evaluate puff enhancement system by speech recognition accuracy. In addition, the performance of puff detection system is evaluated by confusion matrix and ROC curve.

B. Evaluation

1. Speech Recognition Accuracy

The result of speech recognition is shown in Fig. 4. Each sentence and word accuracy is the ratio to recognize whole sentence or partial word. As shown in Fig 4, baseline system improved about 16 %, however, the accuracy of 49.23% is not acceptable in practical recognition system. The proposed system achieved the accuracy of 86.15%.

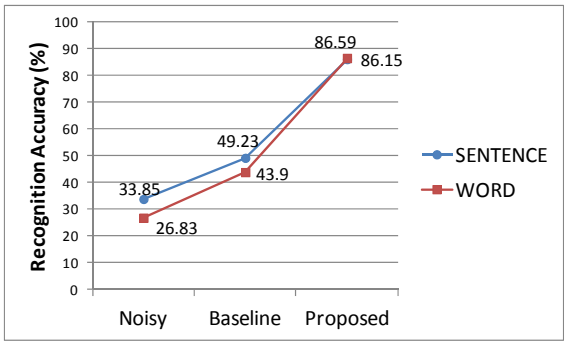

Fig. 4 Speech recognition accuracy

2. Confusion matrix and ROC curve

The confusion matrixes of baseline system and proposed system are shown in Table III and Table IV Both true positive rate and true negative rate of the proposed system are improved compared to those of the baseline system. Fig. 5 represents ROC curve results. The Area under the curve (AUC) of baseline was 0.8408, and that of the proposed system recorded 0.9456 increasing 0.1 from AUC of baseline. It is important that true positive rate was increased because unnecessary filtering in speech interval is reduced. It means that the proposed system has higher reliability.

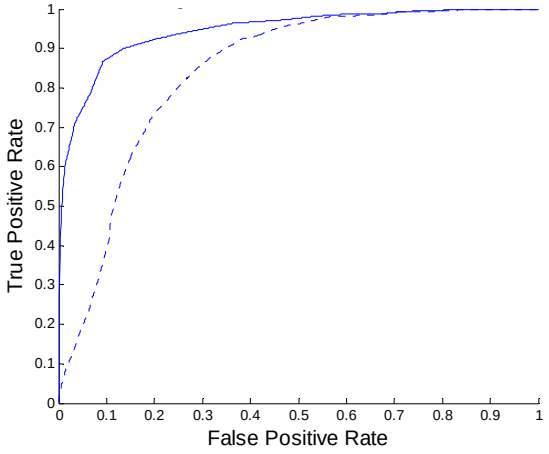

Fig. 5 ROC curve

3. Speech enhancement results

Enhanced speech sample by proposed system is shown in Fig. 6. Noisy sample was misrecognized by concentrated low band energy of puff noise. Proposed enhancement system attenuated the only puff noise interval properly.

V. Conclusion

In this paper, GMM-based puff noise classifier is proposed. The detection system using proposed acoustics features shows improved performance than that of the baseline system using MFCCs. The performance is verified by speech recognition rate after the proposed adaptive high pass filtering. As the results, improvements in speech recognition rate around 37% are achieved.

TABLE III
Confusion Matrix of Baseline System (unit: %)

Confusion Matrix		Predicted Class	
		Clean	Puff
Actual Class	Clean	85.6	14.4
	Puff	39.1	60.9

TABLE IV
Confusion Matrix of Proposed System (unit: %)

Confusion Matrix		Predicted Class	
		Clean	Puff
Actual Class	Clean	94	6
	Puff	22.1	77.9

Acknowledgment

This work was supported by the Industrial Strategic technology development program, 10035252, development of dialog based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Rep. of Korea)

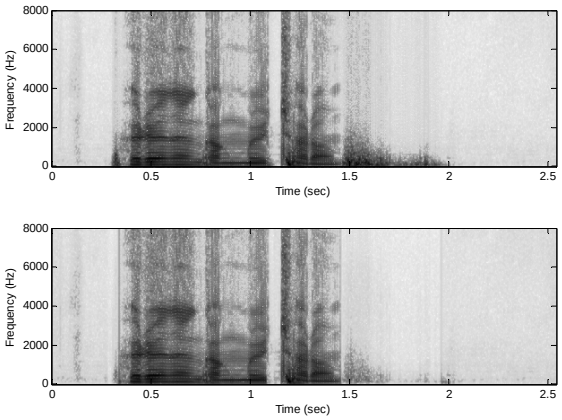

Fig. 6 Spectrogram of enhanced speech
(Top: noisy sample, Bottom: enhanced sample)

References

[1] Pati, V, Rao, P, "Acoustic Features for Detection of Aspirated Stops," *Proc. Of the National Conference on Communications*, 2011
[2] Ishi, C. T, "A New Acoustic Measure for Aspiration Noise Detection", *Proc. of The 8th International Conference of Speech and Language Processing*, 2:941-944, 2004
[3] Schmidt, M., Larsen, J., and Hsiao, F., "Wind noise reduction using non-negative sparse coding", *Proc. of 2007 IEEE Workshop on Machine Learning for Signal*, 2007.
[4] Xiaoqiang, L., Shuangtian, L., Jie, L., "Convolutive Sparse Non-negative Matrix Factorization for Windy Speech", *Proc. of ICSP*, 2010.
[5] Kuroiwa, S., Mori, Y., Tsuge, S., Takashina, M., and Ren, F., "Wind noise reduction method for speech recording using multiple noise templates and observed spectrum fine structure", *Proc. of ICCT*, 2006.
[6] Nemer, E, Leblanc, W, "Single-microphone wind noise reduction by adaptive postfiltering", *Applications of Signal Processing to Audio and Acoustics 2009. WASPAA '09. IEEE Workshop on*., 2009.
[7] Yoshida, M., Oku, T., Yamanaka, M., and Murata, H., "A novel wind noise reduction for digital video camera", *ICCE*, 2008.
[8] Gary, W., Jens, M., Steven, B., and Jurgen, P., "Electronic pop protection for microphones", *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoutics*, 2007.

[9] Lie, L., Hong-Jiang, J., Stan Z., L., "Content-based audio classification and segmentation by using support vector machines", *Multimedia Systems, Springer-Verlag*, 2003
[10] Young, S., et al., "The HTK book (v3.4)", *Cambridge University Press*, 2006.