LBP BASED RECURSIVE AVERAGING FOR BABBLE NOISE REDUCTION APPLIED TO AUTOMATIC SPEECH RECOGNITION

Qiming Zhu and John J. Soraghan

Centre for Excellence in Signal and Image Processing (CeSIP), University of Strathclyde,
Royal College Building, 204 George Street, Glasgow, UK
E-mail: q.zhu@strath.ac.uk, j.soraghan@strath.ac.uk

ABSTRACT

Improved automatic speech recognition (ASR) in babble noise conditions continues to pose major challenges. In this paper, we propose a new local binary pattern (LBP) based speech presence indicator (SPI) to distinguish speech and non-speech components. Babble noise is subsequently estimated using recursive averaging. In the speech enhancement system optimally-modified log-spectral amplitude (OMLSA) uses the estimated noise spectrum obtained from the LBP based recursive averaging (LRA). The performance of the LRA speech enhancement system is compared to the conventional improved minima controlled recursive averaging (IMCRA). Segmental SNR improvements and perceptual evaluations of speech quality (PESQ) scores show that LRA offers superior babble noise reduction compared to the IMCRA system. Hidden Markov model (HMM) based word recognition results show a corresponding improvement.

Index Terms— 1-D LBP, noise estimation, noise reduction, speech recognition, HMM

1. INTRODUCTION

Automatic speech recognition (ASR) is fundamental to a variety of applications such as speech-to-text, speech-to-speech translation, speech command control, speech communication systems. Two of the main challenges that ASR systems must overcome are: 1) obtain the useful information from the speech signal and 2) decrease the effect of noise.

Voice activity detection (VAD) algorithms are designed to detect the speech presence or absence by using speech features. Short-time energy, zero-crossing rate [1] and linear predictive coding coefficients [2] have been used as common features in the early VAD algorithms. In more recent VAD designs, cepstral features [3], formant sharps [4], and least-square periodicity measures [5] were proposed as detection features. The VAD proposed in ITU-T standard G.729 Annex B uses a set of metrics including low-band energy, full-band energy, zero-crossing rate and line spectral frequencies (LSF) to make VAD decision for each 10ms frame of the input signal [6]. Statistical model-based voice activity detection techniques were proposed for noisy speech input [21]. More recently, 1-D LBP was initially proposed to be suit for 1-D signal processing and subsequently applied to onset detection of myoelectric signal [7][8]. It was verified that the 1-D LBP codes are able to distinguish speech presence and absence segments by using the distinguishing LBP codes of higher activity in certain characteristic histogram bins [9].

In ASR systems, speech enhancement systems should be applied before the VAD in order to avoid the effects of noise. Recently, noise estimation based on frequency domain has become popular for speech enhancement. Martin [10] proposed a minimum statistics (MS) algorithm which could estimate the noise power spectrum density. An improved minima controlled recursive averaging (IMCRA) [11] combines MS with recursive averaging to perform the noise estimation was introduced by Cohen. These estimated noise spectra are used as the input to speech enhancement system such OMLSA [12] to obtain higher quality speech signals.

However, experimental results show that IMCRA-OMLSA system cannot effectively reduce the noise in non-stationary babble noise environment. This is due to the speech presence probability (SPP) estimation error in IMCRA process. Unlike our previous works [7][8][9], which applied LBP on the signals themselves, we propose a modified signal energy based LBP calculation for speech presence indicator (SPI). This SPI is combined with recursive averaging to estimate the noise. The OMLSA system enhances the speech signal using this estimated noise. Tests are performed in non-stationary and varying SNR babble noise to show the performance improvement of the LBP based recursive averaging (LRA) system. The segmental SNR (SegSNR) [20] improvements, PESQ [13] scores and hidden Markov model (HMM) based word recognition are compared to the performance obtained from classic IMCRA using real speech signals.

The remainder of the paper is organised as follows. Section 2 describes the algorithms that include LBP based SPI, LBP based recursive averaging combined OMLSA speech enhancement and HMM based speech recognition system. Section 3 provides simulation results and

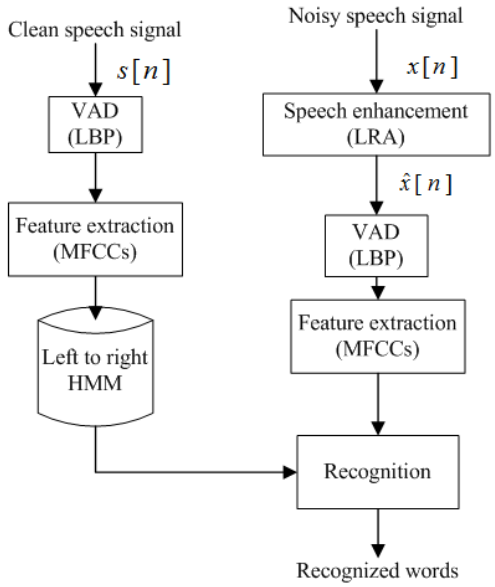

Fig. 1- Speech recognition system

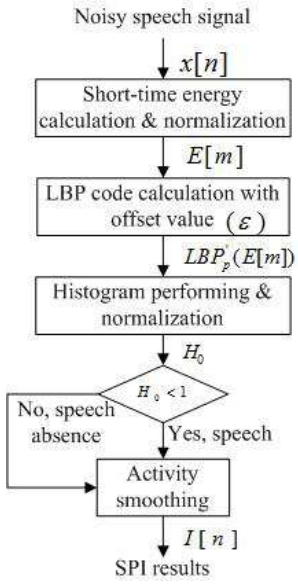

Fig. 2- LBP based SPI

discussions. Conclusions are provided in Section 4.

2. BABBLE NOISE REDUCTION FOR ASR

The overall babble noise reduction and speech recognition system is illustrated in Fig. 1. A LBP based recursive averaging (LRA) process is used in the speech enhancement stage of the system. The authors, in [7][9], have shown that 1-d LBP based VAD is able to efficiently distinguish speech and non-speech segments. It is selected as the VAD in the system.

2.1. Modified LBP based SPI

As introduced by Lamel [14], short-time energy can present the distribution of speech and speech absence segments. Our LBP based SPI is based on thresholding the neighbouring speech signal energy samples of the center energy sample.

As illustrated in Fig. 2, short-time energy $E[m]$ of the noisy signal $x[n]$ is firstly calculated and normalized to be $[0 \rightarrow 1]$. Each energy sample can be presented by a LBP code which is obtained using a new LBP procedure with offset value ε for the energy is defined as follows:

$$LBP'_P(E[m]) = \sum_{r=0}^{\frac{P}{2}-1} \left\{ S\left[E\left[m + r - \frac{P}{2}\right] - E[m] - \varepsilon\right] 2^r + S[E[m+r+1] - E[m] - \varepsilon] 2^{r+\frac{P}{2}} \right\} \quad (1)$$

where P is number of the neighbouring energy samples used. The Sign function $S[\cdot]$ is:

$$S[x] = \begin{cases} 1, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \quad (2)$$

The sample at frame j of histogram $\widehat{\boldsymbol{H}}_{\boldsymbol{0}}$ formed by the LBP code is shown as follows:

$$\hat{H}_0[j] = \sum_{(j-1)\cdot R+1 \leq m \leq j\cdot R} \delta\{LBP'_P(E[m]), 0\} \quad (3)$$

where R denotes the frame size of the histogram, $j = 1,2,\dots,J$. and $\delta(x,y)$ is the Kronecker Delta function. After normalizing $\widehat{\boldsymbol{H}}_{\boldsymbol{0}}$, $\boldsymbol{H_0}$ is then formed that ranges from $[0 \rightarrow 1]$. The detection rule for each frame number j is then applied to $\boldsymbol{H_0}$:

$$\hat{I}[j] = \begin{cases} 1, & H_0[j] < 1 \\ 0, & H_0[j] = 1 \end{cases} \quad (4)$$

$\hat{I}[j]$ is the initial SPI.

An example is shown in Fig. 3. Fig. 3(a) shows typical noisy signal energy with the initial SPI. The gaps, highlighted by the circles in Fig. 3(a) represent the variability in the initial SPI measures. The initial SPI is smoothed using the assumption that the minimum length of the speech segments is set to be 150ms and the maximum gaps length is 80ms i.e.: ignore the short segments and combine the speech segments with short gaps to produce the final SPI $I[j]$. Fig. 3(b) shows the resulting smoothed SPI result. The gaps seen in Fig. 3(a) have been removed to form a smooth SPI function.

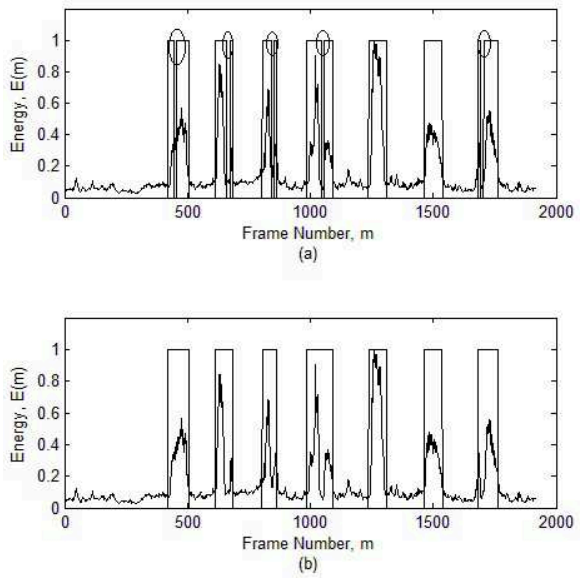

Fig. 3- SPI smoothing example. (a) Noisy signal energy and unsmoothed SPI result. (b) Noisy signal energy and smoothed SPI result

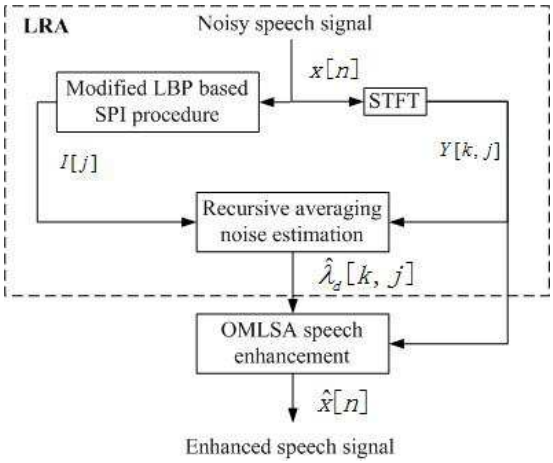

Fig. 4- LRA-OMLSA speech enhancement

2.2 LBP based recursive averaging combined OMLSA (LRA-OMLSA)

A schematic of LBP based recursive averaging combined OMLSA (LRA-OMLSA) speech enhancement unit is shown in Fig. 4. Assume the short-time Fourier transform (STFT) of the noisy signal $x[n]$ is defined as $Y(k,j)$, where k represents the frequency bin index and j is the frame index. As described in [11], the noise spectrum estimator can be denoted as follows:

$$\hat{\lambda}_d(k,j+1) = \beta \cdot \bar{\lambda}_d(k,j+1) \tag{5}$$

where β is a bias compensation vector and $\bar{\lambda}_d$ is the recursive averaging vector which is computed as follows:

$$\bar{\lambda}_d(k,j+1) = \tilde{\alpha}_d(k,j) \cdot \bar{\lambda}_d(k,j) + [1-\tilde{\alpha}_d(k,j)] \cdot |Y(k,j)|^2 \tag{6}$$

where

$$\tilde{\alpha}_d(k,j) \triangleq \alpha_d + (1-\alpha_d) \cdot I(j) \tag{7}$$

where α_d denotes a smoothing parameter, $I[j]$ is the LBP based SPI result. The estimated noise spectrum $\hat{\lambda}_d$ is then used as the input to OMLSA system in order to obtain the enhanced signal $\hat{x}[n]$. Cohen [11] identified the smoothing parameter $\alpha_d = 0.85$ and the bias compensation vector $\beta = 1.47$. In the experiments, the default values of parameters used in OMLSA is the same as described by Cohen [11].

2.3 Recognition system

Our recognition system was illustrated in Fig. 1. Clean speech signals are processed by the VAD stage to separate the speech from non-speech. Mel-frequency cepstral coefficients (MFCCs) of these components are used to train the left-to-right HMMs. The noisy speech signal $x[n]$ is enhanced using the LRA-OMLSA to obtain $\hat{x}[n]$. The MFCCs of the speech components in $\hat{x}[n]$ and the trained HMMs are used in the automatic speech recognition stage.

MFCCs are used to generate the training vectors by transforming the signal into frequency domain. Standard MFCCs computed by the Speech-toolbox [15] are selected as the recognition features.

The HMM Matlab Toolbox [17] is used as the recognition model using the left-to-right HMM [16]. Each word model comprises 6 states, with observations modeled by Gaussian mixture models with 3 components.

The HMM was trained by TI-46 Word Speech Database [18], which contains 46 words spoken by 8 males and 8 females for 10 times (7360 utterances in total). The corpus was recorded at sampling frequency 12.5 kHz. During HMM training, each word has a model which uses LBP based VAD proposed in [7] for speech detection. The recognizer was tested on TI-46 testing dataset recorded by the same speakers 8 times for each word (5888 utterances in total). Babble noise from the NOISEX-92 database are added onto the testing data with SNR range from -5dB to 20dB, in 5dB steps.

TABLE I
LRA COMPARING WITH IMCRA, DESCRIBED BY SEGMENTAL SNR IMPROVEMENTS AND PESQ SCORES

Input SNR level (dB)	SegSNR improvement (dB)		PESQ scores	
	LRA	IMCRA	LRA	IMCRA
10	7.2	3.9	2.88	2.75
8	8.0	4.1	2.75	2.62
6	9.2	4.5	2.64	2.49
4	10.1	4.9	2.47	2.36
2	10.8	5.6	2.41	2.25
0	11.1	6.1	2.29	2.11
-2	10.8	6.5	2.16	1.98
-4	10.2	6.8	2.03	1.85
-6	10.3	7.2	1.94	1.73
-8	10.6	7.7	1.64	1.12
-10	11.5	8.7	1.58	1.05

LRA represent LRA-OMLSA and IMCRA represent IMCRA-OMLSA

3. PERFORMANCE EVALUATION

3.1 Speech enhancement performance

The performance of the LRA-OMLSA for babble noise reduction was tested on 256 speech records from 9 different speaker including 3 females and 6 males. These test speech datasets were obtained from the VoxForge open source [19] at sampling frequency 16 kHz. The clean speech signals are mixed with the non-stationary babble noise from NOISEX-92 database with SNR range from -10dB to 10dB, in 2dB steps. These noisy signals are used for evaluating the speech enhancement systems. Window size for energy calculation is 2.5ms. A value $P = 2$ is used for generating the LBP code that means comparing the energy of neighbouring 160 speech signal samples while calculating LBP code. Offset value $\varepsilon = 0.03$ and the segment size of histogram is 20ms. The values of parameters used in IMCRA-OMLSA are the same as described in [12].

TABLE I shows the SegSNR improvements and PESQ scores for both the LRA-OMLSA and IMCRA-OMLSA speech enhancement. It can be seen that LRA-OMLSA offers a 3dB-5.2dB SegSNR improvement compared to IMCRA-OMLSA. PESQ is well known to correlate highly with mean opinion subjective test scores. It is shown that under all SNR conditions the LRA-OMLSA improves the PESQ results compared to those obtained from the IMCRA-OMLSA. The performance improvement of the proposed method is more profound at low SNR.

3.2 Speech recognition performance

TABLE II shows the recognition rates for different SNR conditions for LRA-OMLSA, IMCRA-OMLSA enhancement and Noisy (no speech enhancement). The results of clean utterances are also compared. Statistically significant best results are in bold. The key findings are as follows:

TABLE II
HMM BASED WORD RECOGNITION RESULTS

SNR level (dB)	LRA	IMCRA	No Speech Enhancement
Clean	**98.23%**	**98.23%**	**98.23%**
20	95.13%	91.88%	**95.47%**
15	**93.91%**	88.86%	90.23%
10	**88.14%**	83.75%	75.71%
5	**81.90%**	69.52%	60.33%
0	**68.37%**	45.42%	31.42%
-5	**57.48%**	38.09%	14.05%

LRA represent LRA-OMLSA, IMCRA represent IMCRA-OMLSA, No Speech Enhancement represent noisy speech signal without speech enhancement.

1) Recognition results without speech enhancement show a relatively poor performance in most noise conditions, suggesting that both 'LRA' and 'IMCRA' significantly decrease the noise effects. However, at relatively low noise conditions the speech enhancer offers no benefit.

2) It is shown that, in all SNR cases, the recognition performance provided by 'LRA' is superior to that obtained using 'IMCRA'. This suggests that LRA combined OMLSA for speech enhancement reduces the babble noise more effectively than the IMCRA-OMLSA algorithm. From the result in Table II, an improvement of approximately of 5.1% at 15dB and 19.4% at -5dB is noted.

Our study confirms that, compared to the IMCRA-OMLSA, LRA-OMLSA for speech enhancement reduces the babble noise more efficiently offering significant speech recognition improvements.

4. CONCLUSION

This paper presented a novel LRA-OMLSA speech enhancement algorithm that was combined with a HMM unit within an automatic speech recognition system. The LRA uses a modified LBP based speech presence indicator (SPI) wherein the histogram of the LBP code is obtained by thresholding the neighbouring energy samples with an offset value ε. LRA estimates the babble noise spectrum which is used as the input to the OMLSA speech enhancement system. By comparing with IMCRA, the SegSNR improvements and PESQ scores showed that the LRA algorithm reduces the babble noise more effectively. Furthermore, HMM based word recognition results show that LRA is superior to IMCRA.

5. REFERENCES

[1] J. C. Junqua, B. Mark, B. Reaves, "A Study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on a DTW and HMM Recognize," Proc. Euro speech 1991, 1371–1374, (1991).
[2] L. R. Rabiner and M. R. Sambur, "Voiced–unvoiced-silence Detection using the Itakura LPC Distance Measure," ICASSP 1977, 323–326, (1977).
[3] J. A. Haigh and J. S. Mason, "Robust Voice Activity Detection using Cepstral Features," in IEEE TEN-CON 1993, 3, 321–324, (1993).
[4] J. D. Hoyt and H. Wechsler, "Detection of Human Speech in Structured Noise," ICASSP 1994, 237–240, (1994).
[5] R. Tucker, "Voice Activity Detection using a Periodicity Measure," IEEE Proc. Com. Speech and Vision, vol. 139(4), 377–380, (1992).
[6] "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70 Annex B," ITU-T Recommendation G.729-Annex B, (1996).
[7] N. Chatlani, J. J. Soraghan, "Local Binary Patterns for 1-D Signal Processing", EUSIPCO 2010, 95-99, (2010).
[8] P. McCool, N. Chatlani, L. Petropoulakis, J. J. Soraghan, R. Menon and H. Lakany, "1-D Local Binary Patterns for Onset Detection of Myoelectric Signals," EUSIPCO 2012, 1633–1637, (2012).
[9] Q. Zhu, N. Chatlani and J. J. Soraghan, "1-D Local Binary Patterns Based VAD using in HMM-based Improved Speech Recognition," EUSIPCO 2012, 1633–1637, (2012).
[10] R. Martin, "Noise PSD estimation based on optimal smoothing and minimum statistics," IEEE Trans. Audio, Speech Lang. Process, 9(5), 504-512, (2001).
[11] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," IEEE Trans. Audio, Speech Lang. Process, 11(5), 466-475, (2003).
[12] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Processing. Amsterdam, The Netherlands: Elsevier, 81, 2403–2418, (2001).
[13] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Rec. P.862, (2001).
[14] L. Lamel, L. Rabiner, "An improved endpoint detector for isolated word recognition," IEEE Trans. Audio, Speech Lang. Process, 29(4), 777-785, (1991).
[15] Q. He, Y. He, "Matlab extended," Tsinghua Uni. Proc., (2002).
[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, 77(2), (1989).
[17] K. Murrphy, "Hidden Markov model (HMM) toolbox for Matlab," Online: http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html (Last viewed Oct. 23, 2013).
[18] M. Liberman, et al., "TI 46 word speech database: speaker-dependent isolated word corpus," Online: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC93S9 (Last viewed Oct. 23, 2013)
[19] VoxForge Speech Corpus, available: http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/ (Last viewed Oct. 25, 2013)
[20] S. Quackenbush, T. Barnwell and M. Clements, "Objective Measures of Speech Quality", Prentice-Hall, *Englewood CliPs*, NJ, 1988.
[21] J. Sohn, N. S. Kim, and W. Sung. "A statistical model-based voice activity detection." *IEEE Signal Processing Letter*, 6 (1): 1–3, 1999.