LOCAL PEAK ENHANCEMENT COMBINED WITH NOISE REDUCTION ALGORITHMS FOR ROBUST AUTOMATIC SPEECH RECOGNITION IN AUTOMOBILES

Osamu Ichikawa, Takashi Fukuda, and Masafumi Nishimura

Tokyo Research Laboratory, IBM Japan Ltd., Yamato-shi, Kanagawa-ken, Japan, 242-8502
{ICHIKAW, FUKUDA1, NISIMURA}@jp.ibm.com

ABSTRACT

The accuracy of automatic speech recognition in automobiles is significantly degraded in very low SNR (Signal to Noise Ratio) situations such as "Fan high" or "Window open". In such cases, speech signals are often buried in broadband noise. In this paper, we propose a novel approach for such situations that utilizes harmonic structures in the human voice. It pursues two objectives. (1) Unlike comb filtering, it should not rely on F0 detection or voiced/unvoiced detection, since they are not accurate enough in noisy environments. (2) It should work with existing noise reduction algorithms. In our new approach, an observed power spectrum is directly converted into a filter for speech enhancement by retaining only the local peaks considered to be harmonic structures. In our experiments, we reduced the word error rate significantly in realistic automobile environments, and our approach showed further improvements when used with existing noise reduction algorithms.

Index Terms— Harmonic analysis, Speech enhancement, Speech recognition, Noise, Robustness

1. INTRODUCTION

The performance of automatic speech recognition in automobiles is affected by various noises. Beamformer [1] technology reduces directional noise such as voices from passengers and sounds coming from a car radio, TV, or CD player. However, it does not have sufficient signal recovery in very low SNR situations with ambient noise (such as "Fan high" or "Window open") unless the size of the beamformer is very large. For single channel signal processing, existing noise reduction algorithms such as a Wiener Filter [2] or Spectral Subtraction (SS) [3] are known to improve the accuracy, but improvements are still needed in those situations. Therefore, different approaches beyond reducing noise should be combined with existing noise reduction algorithms.

One of the candidate approaches involves enhancements of the harmonic structures in human voices. Comb filtering [4] and its variants [5] were proposed and showed good performance, especially in mixed speech cases. However, they are rarely integrated into commercial ASR products, and especially not for automobiles. This is because designing a comb filter relies on the accurate estimation of F0 (the fundamental frequency) and the accurate discrimination between voiced and unvoiced speech. It was reported that errors at this stage have detrimental effects on the performance [6]. Szymanski et al. proposed Comb Filter Decomposition [7] that does not require F0 estimation, but their experiment was limited to white Gaussian noise.

Another candidate would use a matching algorithm to put larger weights on frequencies having larger spectral powers as the decoder calculates likelihoods [8][9]. This is based on the assumption that frequencies having more spectral power are noise robust and most likely to be the formant frequencies in voiced speech frames. Huang et al. enhanced the logic for the MFCC domain [10], but this involved adding autocorrelation into their decoding process.

In this paper, we propose a novel approach for the speech enhancement. It uses a filter designed to enhance the harmonic structure which is observed as local peaks at regular distances in the spectrum domain. It does not depend on F0 or voiced/unvoiced detection. Since it works as a front-end for both training and decoding, it does not require any changes in existing decoders. This new method will be referred to as LPE (Local Peak Enhancement) in the following sections.

2. PROPOSED METHOD

2.1. LPE

Fig. 1 shows the whole process of LPE and sample outputs at each step for both a voiced frame and a noise frame. The process is the same for entire frames, but the generated filter looks very different depending on whether or not the frame is voiced speech, as shown in the figure.

In the first step, an observed spectrum $y_T(j)$ is converted to a log power spectrum $Y_T(j)$.

$$Y_T(j) = \log(y_T(j)) \tag{1}$$

Here, the index T is a frame number and j is the bin number of the DFT corresponding to the subband frequency. The process described in this section should be performed for each T.

Then the log power spectrum is converted to a cepstrum $C_T(i)$ by using $D(i,j)$, a DCT (Discrete Cosine Transformation) matrix.

$$C_T(i) = \sum_j D(i,j) \cdot Y_T(j) \tag{2}$$

The cepstra represent the curvatures of the log power spectra. The lower cepstra correspond to long oscillations, and the upper cepstra correspond to short oscillations. We need only the medium oscillations. The range of the cepstra is chosen to cover possible harmonic structures in the human voice. Therefore the lower and the upper cepstra should be filtered out.

$$\begin{aligned} \hat{C}_T(i) &= epsilon \cdot C_T(i) && \text{if } i < lower\_cep \text{ or } i > upper\_cep \\ &= C_T(i) && \text{else} \end{aligned} \tag{3}$$

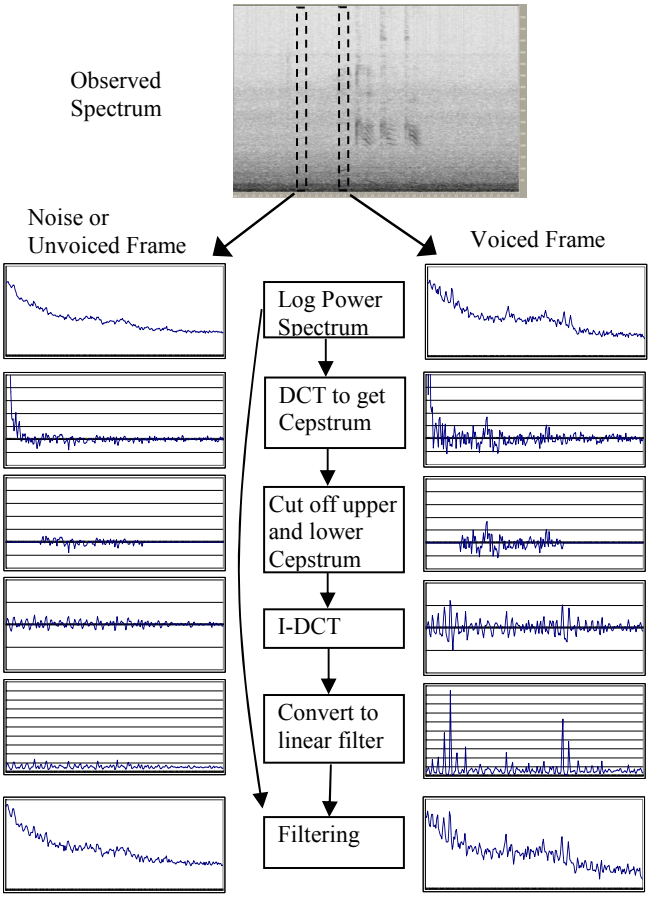


Fig. 1. Process of LPE

In our experiments, *lower_cep*=40 and *upper_cep*=160 for a 16 KHz sampling frequency with an FFT length of 512 samples. This corresponds to an F0 range from 100 Hz to 400 Hz for the human voice, with *epsilon* being close to zero. We set it to 10^{-3}.

The filtered cepstrum $\hat{C}_T(i)$ is converted back to a log power spectrum by using an I-DCT.

$$W_T(j) = \sum_i D^{-1}(j,i)\cdot\hat{C}_T(i) \tag{4}$$

Then it is converted back to a linear power spectrum, and it is normalized so that the average is 1.0.

$$w_T(j) = \exp(W_T(j)) \tag{5}$$

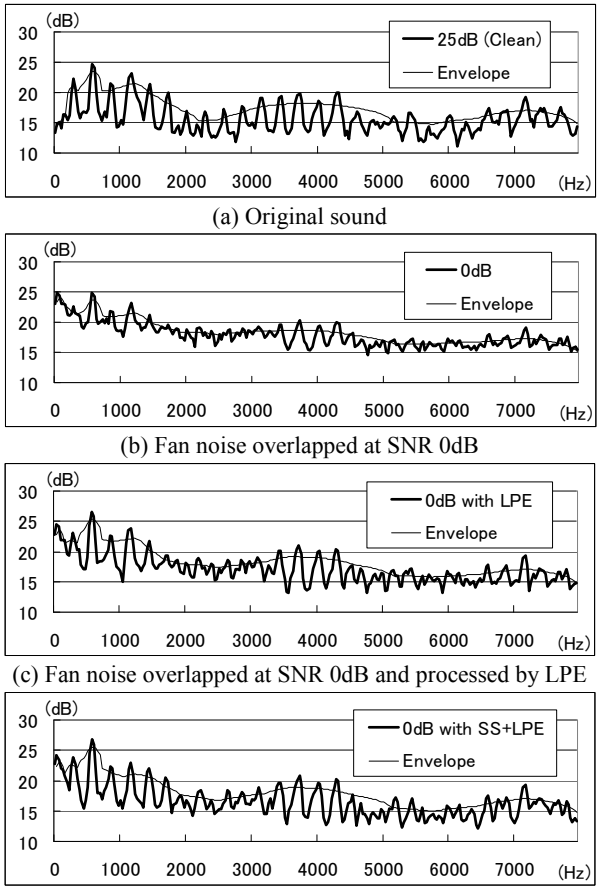
$$\overline{w}_T(j) = w_T(j)\cdot\frac{Num\_bin}{\sum_k w_T(k)} \tag{6}$$

Here, *Num_bin* is the number of bins used in the FFT. The filter is obtained as $\overline{w}_T(j)$. Finally, the enhanced output $z_T(j)$ is obtained as

$$z_T(j) = \overline{w}_T(j)\cdot y_T(j) \tag{7}$$

In order to reduce the amount of computation, the steps of the Equations (2), (3), and (4) can be combined into a single step using the pre-calculated matrix A as follows.

$$\begin{aligned}\Lambda(i,j) &= 0 && \text{if } i \neq j\\ &= epsilon && \text{else if i < lower\_cep or i > upper\_cep}\\ &= 1 && \text{else}\end{aligned} \tag{8}$$



(a) Original sound

(b) Fan noise overlapped at SNR 0dB

(c) Fan noise overlapped at SNR 0dB and processed by LPE

(d) Fan noise overlapped at SNR 0dB and processed by LPE after SS

Fig. 2. Spectrums of vowel /u/ recorded in a stationary car with and without fan noise overlapping at the specified SNR. The spectrum envelope is plotted with Mel-Filtering.

$$A = D^{-1}\Lambda D \tag{9}$$

$$W_T = AY_T \tag{10}$$

2.2. Characteristics of an LPE Filter

As shown in Fig. 1, the filter for LPE is derived directly from the observed spectrum. Therefore, F0 estimation is not required. For a noise frame or an unvoiced speech frame, it will be designed to be almost flat. This means LPE does almost nothing to such frames, and therefore, LPE does not require voiced/unvoiced detection.

For voiced speech frames, the LPE filter is designed to enhance the harmonic structures in the observed spectrum. Unlike a comb filter, the LPE filter is not uniform over all frequencies. It is more focused on the frequencies where harmonic structures are observed in the input spectrum. Therefore the acoustic model should be retrained with LPE for automatic speech recognition.

Fig. 2 shows how a spectrum is degraded by a noise. In Fig. 2(a), the original clean spectrum shows three formants around 600 Hz, 1200 Hz, and 3500 Hz. However, in Fig. 2(b), they are less conspicuous, and the spectrum contour is close to flat. In contrast, LPE retains more of the characteristics of the formants, as shown

in Fig. 2(c). The combination of SS and LPE retains even more, as shown in Fig. 2(d). An advantage of LPE is that voiced speech immersed in heavy noise should be more distinct and distinguishable for decoding.

Harmonic structures are conspicuous around frequencies having larger spectral powers in the voiced speech frames, and they are most likely to be formant frequencies. Therefore, this approach inherently involves formant enhancement as well as harmonic enhancement, under the assumption that the noise has a broad spectrum and the harmonic structure is not locally destroyed by the noise.

3. EXPERIMENTS

CENSREC-3, a common evaluation framework for isolated Japanese word recognition in actual automobile environments was used in this experiment. This data was collected by IPSJ, and is widely used to evaluate noise reduction algorithms [11]. It has speech data both for training and testing for automatic speech recognition using multi-style trained acoustic models.

The test data in the database was recorded under 16 environmental conditions using combinations of three vehicle speeds and six kinds of in-car environments as shown in Table 1. A total of 14,216 utterances spoken by 18 speakers (8 males and 10 females) were recorded at a 16 KHz sampling frequency. The recognition grammar is a list of 50 words.

For training, each driver's speech saying phonetically balanced sentences was recorded under two conditions: while idling and while driving on a city street in a normal in-car environment. A total of 14,050 utterances spoken by 293 drivers (202 males and 91 females) were recorded with a close-talking microphone and a hands-free microphone.

In this experiment, we used only hands-free microphone data for both training and testing. The acoustic models were trained with both idling data and driving data for the front-end processing being tested. This corresponds to Condition 3 as defined in CENSREC-3. The evaluation category is zero, which means no changes at the backend using HTK with 39 dimensional features (12 mel-cepstrum + log power, with their Δ and $\Delta\,\Delta$) without subtracting cepstrum mean.

Comb filtering was introduced as a conventional method to compare. It uses F0 estimation and voiced/unvoiced detection. We used the "Pitch command" in SPTK-3.0 [12] to obtain this information. We used a low-end frequency of 100 Hz and an upper frequency limit of 400 Hz, so to be compatible with the LPE experiments. The voiced/unvoiced threshold was empirically set to 7.0, because this gave us better results than the SPTK default value.

Table 1 shows the resulting word accuracies for various environmental conditions. The baseline is the evaluation without using any speech enhancement or noise reduction algorithms. Table 1 also shows the estimated SNRs of the test data using the VAD (Voice Activity Detection) information from the ETSI Advanced Front-End (ES202-050) [2]. Note that the accuracy of the SNR depends on the VAD information. Table 2 shows the estimated SNRs of the training data. We see CENSREC-3 trains an acoustic model at relatively better SNRs than for the test data. Therefore, speech enhancement and noise reduction are expected to help the test performance.

LPE enhances the local peaks considered to be harmonic structures. Therefore, a drawback is expected with LPE when the background noise contains music or speech from audio devices

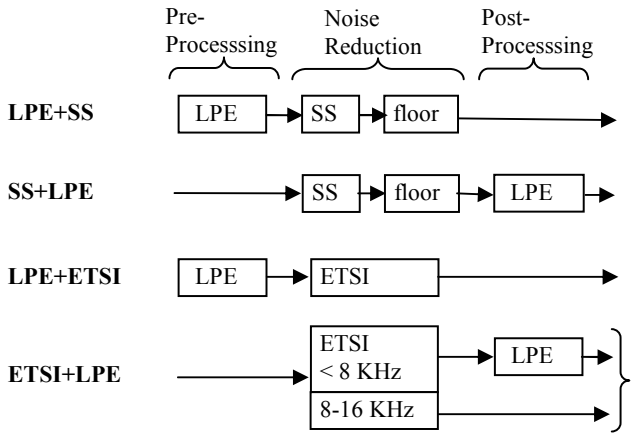


Fig. 3. Combinations of LPE and noise reduction algorithms

such as a radio, TV, or CD player, because the filter is designed to enhance that audio, too. This is a known restriction of LPE. Comb filtering shares this problem, and a multi-pitch tracker was proposed to address it [13]. In this paper, we accept this restriction and we focus only on the results of the "Audio off" cases. The restriction should not matter with current car navigation systems, because most of them are designed to disable audio on pushing a talk button. Also, we can expect an echo canceller to eliminate audio components before processing by LPE.

For the average "Audio off" case, LPE outperformed the baseline by 17.0% in error reduction. Most of the improvement was gained in the very noisy conditions of "Fan high" and "Window open" conditions with error reductions of 14.8% and 23.7%, respectively. Comb filtering also improved the accuracy in these conditions. However, the improvement was smaller than LPE.

In relatively clean conditions such as "Normal" or "Fan low" at "Idling" or "Low speed", the accuracy of LPE was almost the same or slightly degraded from the baseline. However, the degree of loss was small enough for practical use. In contrast, comb-filtering shows noticeable degradation in these conditions. As the "Pitch command" in SPTK works on a per-frame basis, this result could be improved by using a frame-tracking algorithm.

LPE can be used in combination with existing noise reduction algorithms. SS and ETSI ES202-050 were used in our evaluations. For the SS processing, the first 0.1 sec. of each utterance was assumed to be a non-speech segment where the noise spectrum could be estimated. The subtraction weight was set to 1.0, and the flooring coefficient was set to 0.1. As shown in Fig. 3, "LPE+SS" means LPE pre-processes the input of SS, and "SS+LPE" means LPE post-processes the output of SS. Since ETSI ES202-050 splits the 16-KHz input into a less-than-8-KHz part and an upper-8-KHz part, "ETSI+LPE" applied LPE only to the less-than-8-KHz part of the ETSI ES202-050 output. All of the combinations outperformed the noise reduction algorithm or LPE alone in the average "Audio off" case. The pre-processing case and the post-processing case performed almost the same in the "Audio off" cases, but we recommend the post-processing combinations, because the degradations in the "Audio on" cases were smaller. In this evaluation, the best combination was "ETSI+LPE", which reduced the error rate by 69.2% from the baseline in the average "Audio off" case.

Table 1. Word accuracy and estimated SNRs for various environmental conditions. SNR was calculated for the baseline data after a 250 Hz high-pass filtering.

CENSREC-3 (Condition 3)			SNR (dB)	Word Accuracy (%) Base-line	Comb Filter	LPE	Word Accuracy (%) SS	LPE + SS	SS + LPE	ETSI	LPE + ETSI	ETSI + LPE
Idling	Audio off	Normal	16.2	99.7	98.8	99.7	99.8	99.6	99.0	100.0	99.8	100.0
		Hazard on	15.3	98.7	95.3	96.8	96.8	96.9	96.7	98.1	98.1	98.6
		Fan low	11.3	94.6	87.7	94.8	95.2	95.7	95.3	99.2	99.6	99.7
		Fan high	6.2	53.4	55.0	60.3	58.1	65.7	67.6	85.3	89.9	88.9
		Window open	10.5	90.0	85.4	92.7	90.4	94.1	93.8	97.2	98.2	98.0
	Audio on		9.9	81.4	73.2	56.4	74.8	57.0	61.4	89.5	77.7	82.6
Low speed	Audio off	Normal	10.9	99.3	96.6	98.7	98.4	97.8	97.5	99.7	98.6	99.7
		Fan low	9.7	95.1	91.8	94.7	94.6	94.4	94.2	97.8	97.5	98.7
		Fan high	6.7	62.7	66.2	69.1	66.9	71.1	74.3	87.9	89.5	91.5
		Window open	9.3	66.2	70.6	74.3	72.4	76.7	78.5	87.0	89.6	88.7
	Audio on		6.7	79.0	74.7	61.6	79.5	62.1	62.8	90.8	81.3	87.6
High speed	Audio off	Normal	7.5	95.0	94.3	96.2	97.8	95.3	95.9	98.1	97.2	98.8
		Fan low	7.1	89.0	86.7	89.7	91.7	91.9	91.6	96.7	94.8	97.6
		Fan high	6.1	58.2	62.1	63.6	61.3	68.3	69.6	88.4	89.1	88.1
		Window open	7.2	22.2	35.8	40.4	40.1	44.2	45.4	65.0	69.4	66.7
	Audio on		3.9	79.3	69.0	66.6	84.3	67.4	69.1	92.8	84.0	89.7
Average (ALL)				78.9	77.6	78.4	81.3	79.8	80.7	92.1	90.9	92.1
Average (Audio off)				**78.8**	**78.9**	**82.4**	**81.8**	**84.0**	**84.6**	**92.3**	**93.2**	**93.5**
Average (Audio on)				79.9	72.3	61.5	79.5	62.2	64.4	91.0	81.0	86.6
Average (Fan high)				58.1	61.1	64.3	62.1	68.4	70.5	87.2	89.5	89.5
Average (Window open)				59.5	63.9	69.1	67.6	71.7	72.6	83.1	85.7	84.5

4. CONCLUSION

We proposed a novel approach to speech enhancement to improve automatic speech recognition in very noisy conditions. It generates a filter to enhance the harmonic structures observed in the input spectrum, without relying on F0 estimation and voiced/unvoiced detection. Experiments using automatic speech recognition showed this method significantly improved the accuracy in very noisy conditions such as "Fan high" or "Window open". We also confirmed this method can be a pre-processor or a post-processor of existing noise reduction algorithms such as SS and ETSI ES202-050 for further improvements. The drawbacks in "Audio on" cases were smaller in post-processing cases than in pre-processing cases.

Table 2. Estimated SNRs of CENSREC-3 training data.

Training Data	SNR (dB)
Idling	21.1
Driving	18.7

REFERENCES

[1] H. Saruwatari, K. Sawai, A. Lee, K. Shikano, A. Kaminuma, and M. Sakata, "Speech enhancement and recognition in car environment using blind source separation and subband elimination processing", *Proc. of 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 367-372, 2003.

[2] ETSI ES 202 050 v1.1.1, "Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms", 2002.

[3] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech & Signal Process.*, Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

[4] H. Tolba and D. O'Shaughnessy, "Robust automatic continuous-speech recognition based on a voiced-unvoiced decision", *Proc. of ICSLP*, paper 0342, 1998.

[5] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment", *Proc. of ICASSP*, Vol. 1, pp. 125-128, 2001.

[6] T. Nakatani, T. Irino, and P. Zolfaghari, "Dominance spectrum based V/UV classification and F0 estimation", *Proc. of EuroSpeech*, pp. 2313-2316, 2003.

[7] L. Szymanski and M. Bouchard, "Comb filter decomposition for robust ASR", *Proc. of InterSpeech*, pp. 2645-2648, 2005.

[8] M. Sugiyama and K. Shikano, "LPC peak weighted spectral matching measures", *ASJ Trans. of the Com. on Speech Res.*, S80-13, pp. 101-108, 1980.

[9] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features", *Acoustical Society of America Journal*, Vol.116, Issue 4, pp. 2480-2480, 2004.

[10] C. Huang, Y. Huang, F. Soong, and J. Zhou, "Weighted likelihood ratio (WLR) hidden Markov model for noisy speech recognition", *Proc. of ICASSP*, Vol. 1, 2006.

[11] M. Fujimoto, et al., "CENSREC-3: Data collection for in-car speech recognition and its common evaluation framework", *Proc. of International Workshop on Real-world Multimedia Corpora in Mobile Environments*, RWCinME2005, pp. 53-60, 2005.

[12] http://www.sp.nitech.ac.jp/~tokuda/SPTK/

[13] M. Wu, D. Wang, and G.J. Brown, "A multi-pitch tracking algorithm for noisy speech", *Proc. of ICASSP*, Vol. 1, pp. 369-372, 2002.