Speech enhancement in noisy environments for video retrieval

Huiyu Zhou, Abdul Sadka and Richard M. Jiang
Brunel University, Uxbridge, Middlesex, United Kingdom
E-mail:{Huiyu.Zhou, Abdul.Sadka, Min.Jiang@brunel.ac.uk}

Abstract

In this paper, we propose a novel spectral subtraction approach for speech enhancement via maximum likelihood estimate (MLE). This scheme attempts to simulate the probability distribution of useful speech signals and hence maximally reduce the noise. To evaluate the quality of speech enhancement, we extract cepstral features from the enhanced signals, and then apply them to a dynamic time warping framework for similarity check between the clean and filtered signals. The performance of the proposed enhancement method is compared to that of other classical techniques. The entire framework does not assume any model for the background noise and does not require any noise training data.

1. Introduction

Automatic indexing and retrieval of audiovisual data has vital applications in professional media production, media archive management, education, and surveillance. Nowadays, there are a significant number of footages that have been shot but not ever used [1]. These footages, normally, have not been properly indexed. As the volume of the available media information becomes exponentially increasing, manual indexing and retrieval is almost impossible. One of the most proper ways to achieve automatic indexing and retrieval is through content based video analysis of audiovisual data.

Current approaches for video indexing are mainly focused on visual information, e.g. color histograms [5], motion vectors [11] and key frames. The contributions from the content of the accompanying audio signals has not attracted sufficient attention. In fact, to some extent critical video information can be better represented in the continuous flow of audio signals rather than the pictorial part. For instance, gun fighting in a video scene can be much easily detected using sound measurements, while the image content may vary significantly from one frame to another. Another example is that in some movies image shots can be "frozen" for a while but the accompanying music/speech/background sound indicates the continuous progress of different events.

Recently, integration of audio and visual information within a single framework has demonstrated its effectiveness in video retrieval, e.g. [9]. Audio signals, as a counterpart of visual information, have been used to segment scenarios in video sequences due to semantic content discrimination capabilities. In addition, audio source parsing and indexing aids the extraction of a speaker label mapping of the source over time. Integration of the audio and visual mappings constrained by interaction rules hence leads to high-level video abstraction and partial detection of its context [10]. Similar work was also reported in [8].

It has been well recognised that extraction of auditory features plays a key role in an audiovisual retrieval system. Such a feature extraction aspect has been overwhelmingly studied in noise-free or weakly noisy environments. Nevertheless, audio signals with intensive background noise commonly appear in video sequences, leading to degenerate performance of classical sound detection/recognition algorithms. Therefore, appropriate audio enhancement needs to be conducted so as to minimise the effects of background noise on audio detection/recognition. Traditional audio enhancement have been established using expectation-maximisation algorithm based Gaussian Mixture Models scheme [12], non-linear spectral subtraction [2], multi-band spectral subtraction [7] and log-spectral minimum mean square error method [4].

In this paper, we intend to explore an optimal sound enhancement algorithm. Particularly, we are mainly interested in human speech collected in noisy situations. Our method is illustrated in Fig. 1. We propose a novel spectral subtraction approach for speech enhancement via maximum likelihood estimate, namely "MLESS". This scheme attempts to simulate the probability distribution of useful speech signals and hence maximally reduce the noise in spectral domain. To evaluate the quality of speech enhancement, we extract cepstral features from the enhanced signals, and then apply them to a dynamic time warping framework for similarity check between the clean and filtered signals. The higher

978-0-7695-3130-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WIAMIS.2008.38

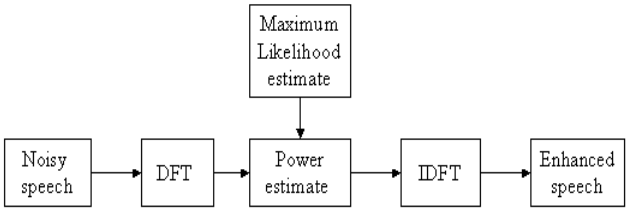

Figure 1. Flowchart of the proposed speech enhancement algorithm.

similarity, the better enhancement quality. This proposed speech enhancement method can be easily encoded into an audiovisual retrieval system in the future.

2. Spectral subtraction by maximum likelihood estimation

2.1 Noise reduction via spectral subtraction

Suppose speech signals $x(m)$ is contaminated by background noise $n(m)$. Then the speech samples can be expressed as

$$y(m) = x(m) + n(m). \quad (1)$$

In the frequency domain, this may be denoted as: $Y(j\omega) = X(j\omega) + N(j\omega)$, where $Y(j\omega)$, $X(j\omega)$, $N(j\omega)$ are Fourier transforms of $y(m)$, $x(m)$, and $n(m)$, respectively.

The statistic parameters of the noise cannot be known, so the estimates of the noise and speech signals are represented by $\hat{X}(j\omega) = Y(j\omega) - \hat{N}(j\omega)$. Normally, the noise spectrum estimate $\hat{N}(j\omega)$ is calculated using the time averaged noise spectrum: $\hat{N}(j\omega) = \mathbf{E}[|N(j\omega)|] \cong |\bar{N}(j\omega)| = \frac{1}{k}\sum_{i=0}^{k-1}|N_i(j\omega)|$, where $|\bar{N}(j\omega)|$ is the amplitude spectrum of the i-th of the k frames of noise. Using truncated Fourier series, we have $\hat{N}_k(j\omega) = |\tilde{N}_k(j\omega)| = \lambda_n|\tilde{N}_{k-1}(j\omega)| + (1-\lambda_n)|\tilde{N}_{k-1}(j\omega)|$, where $\tilde{N}_k(j\omega)$ is the estimate of the filtered noise at i-th frame, λ_n is the filtering coefficient ($0.5 \leq \lambda_n \leq 0.9$). To achieve the noise estimate, the part of the original sound only with noise must be independently analysed.

The spectral subtraction error can be defined as:

$$\epsilon = \hat{X}(j\omega) - X(j\omega). \quad (2)$$

This error can be approximated as $\epsilon \cong |\bar{N}(j\omega)| - \mathbf{E}[|N(j\omega)|]$. The signal-to-noise ratio (SNR) can be defined in frequency domain as SNR^c (for clean signals) or SNR^n (for noisy signals). They are represented by: $SNR_k^c(\omega) = |X_k(j\omega)|^2/|N_k(j\omega)|^2$, and $SNR_k^n(\omega) = |Y_k(j\omega)|^2/|N_k(j\omega)|^2$. In the restoration process, the clean signal is unknown, and hence the SNR needs to be estimated. Using the Gaussian model, an optimal SNR at k-th frame can be defined as: $snr_k^c(\omega) = (1-\beta)P(SNR_k^n(\omega) - 1) + \beta|\hat{X}_{k-1}(j\omega)|^2/|\hat{N}_k(j\omega)|^2$, where $SNR_k^n(\omega) = |\hat{Y}_k(j\omega)|^2/|\hat{N}_k(j\omega)|^2$ and $P(x) = \begin{cases} x; (x \geq 0) \\ 0; (otherwise) \end{cases}$. To reduce the level of noise a non-linear spectral subtraction method was devised to over-subtract the noise, based on the signal-to-noise ratio [2]. $|D(j\omega)|^2 = |Y(j\omega)|^2 - \alpha\mathbf{E}|N(j\omega)|^2$, and $|X(j\omega)| = \begin{cases} |D(j\omega)|; (|D(j\omega)| > \theta|N(j\omega)|) \\ \theta|N(j\omega)|; (otherwise) \end{cases}$ with the subtraction factor $\alpha \geq 1$ and $\alpha = \alpha_0 - SNR/s$ (α_0 is the initial value at $SNR = 0$, $1/s$ is the slope); the spectral floor parameter θ is constrained by $0 < \theta \ll 1$.

2.2 Maximum likelihood estimation

From the introduction in the previous subsection, one can easily discover that classical non-linear spectral subtraction methods mainly rely on heuristic parameters (e.g. α and θ) so as to optimally subtract noise from the signals. They have had promising performance with optimal noise reduction, especially eliminating the annoying "musical noise". However, due to the fixed parameters the speech enhancement cannot be properly adapted to generic situations of different noise properties. Thus, these methods have to be improved in order to keep their performance in these circumstances. We here propose a new approach based on maximum likelihood estimation.

Consider a dynamic representation for the speech enhancement, $f(\phi 1_i, \phi 2_i, \phi 3_i)$, where $\phi 1_i$ is the spectrum of the audio signal, $\phi 2_i$ is the spectrum of the enhanced data, and $\phi 3_i$ is the enhancement model at time i. Given a proper $\phi 3_i$, we can obtain optimal $\phi 2_i$ from $\phi 1_i$ using maximum likelihood. In other words, we pursue a maximised $p(\phi 2_i|\phi 3_i)$.

Assuming individual audio samples are conditionally independent, the joint probability is therefore $p(\phi 2_i|\phi 3_i) = \prod_j p(\phi 2_{ij}|\phi 3_i)$, where j is the index of one of the audio samples. Using Bayes' rule, $p(\phi 2_{ij}|\phi 3_i) = \sum_l p(\phi 2_i|\phi_{1l})\phi 3_i)$, where l is also a sample under investigation. The maximum log-likelihood estimate of $\phi 3_i)$ can be defined as $\mathcal{L}(\phi 3_i) \equiv \log p(\phi 2_i|\phi 3_i)$. To achieve this target, we utilise an expectation-maximisation (EM) algorithm.

2.3 EM-based speech enhancement

The EM algorithm starts from an initial signal-to-noise ratio computation, based on the heuristic parameters used in the classical non-linear spectral subtraction methods.

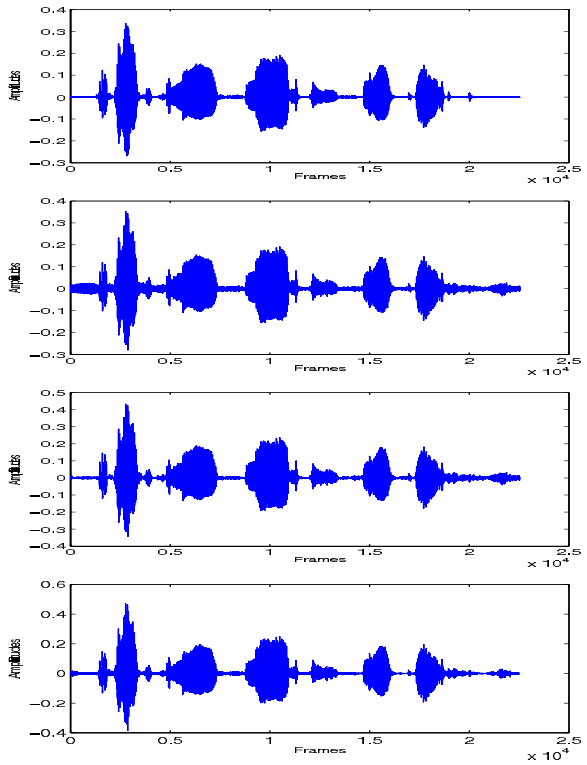

Figure 2. Performance illustration of speech enhancement algorithms: 1st row - clean signal; 2nd row - clean signal with real noise of 15dB (SNR); 3rd row - classical non-linear spectral subtraction [2]; final row - proposed EM-based algorithm.

Then the following E- and M-steps are iterated via the expectation-maximisation algorithm [3]. In fact, a neighborhood around each heuristic parameter is searched for exploring a maximum SNR value. The searching rule is described as follows:

(1) E-stage: we formulate the conditional probability of the enhanced signals to be $p(\phi 1_{il}|\phi 2_{ij}, \phi 3_{i-1})$, which can be expressed as: $p(\phi 1_{il}|\phi 2_{ij}, \phi 3_{i-1}) = \frac{p(\phi 1_{il}|\phi 3_{i-1})p(\phi 2_{ij}|\phi 1_{il},\phi 3_{i-1})}{\sum_m [p(\phi 1_{il}|\phi 3_{i-1})p(\phi 2_{ij}|\phi 1_{il},\phi 3_{i-1})]}$. Hence, we have another form of maximum log-likelihood estimate as follows: $Q(\phi 3_i|\phi 3_{i-1}) \propto \sum\sum p(\phi 1_{il}|\phi 2_{ij}, \phi 3_{i-1})(\phi 1_{il} - \phi 2_{ij})(\phi 1_{il} - \phi 2_{ij})^T$.

(2) M-stage: we expect to obtain the maximum log-likelihood estimate so that $Q(\phi 3_i|\phi 3_{i-1}) \geq Q(\phi 3_{i-1}|\phi 3_{i-1})$. This E-M iteration will terminate if and only if $|Q(\phi 3_i|\phi 3_{i-1}) - Q(\phi 3_{i-1}|\phi 3_{i-1})|$ is less than a threshold.

Fig. 2 illustrates the performance comparisons of the

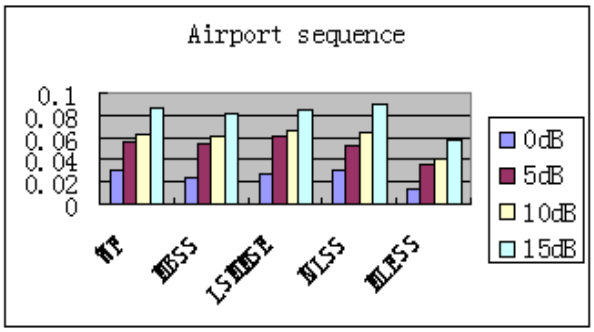

Figure 3. Performance comparison of different speech enhancement methods in the "airport" sequence.

classical non-linear spectral subtraction and the proposed EM-based algorithm. It is observed that the proposed algorithm has subjectively led to better speech restoration.

3. Experimental work

In this section, we evaluate the proposed speech enhancement algorithm by comparing its performance with that of Wiener filtering (WF), multi-band spectral subtraction (MBSS) [7], log-spectral MMSE (LSMMSE) [4], non-linear spectral subtraction (NLSS) [2]. Here, we apply a dynamic time warping based similarity check between the noise-free signals and the noisy signals to be restored. Dynamic time warping is used for similarity check due to its capability in handling two sequences which may vary in time or speed. To conduct reliable similarity check, cepstral features need to be extracted, resulting from effective reduction of environmental sound using silence chopping, emphasizing and segmentation schemes.

We use a noisy speech corpus (NOIZEUS) that contains 30 IEEE sentences corrupted four different real world noises at different SNRs. The noise signals were added to the speech signals at SNRs of 0dB, 5dB, 10dB, and 15dB. The sentences were originally sampled at 25 kHz and down-sampled to 8 kHz. The noise was taken from the AURORA database and includes suburban train noise, airport, street and car noise [6].

For example, Fig. 3 illustrates the performance comparison of different speech enhancement methods in the "airport" sequence. The values shown in the figure indicate the distance metrics. Therefore, smaller values correspond to better similarity (enhancement quality) between the noise-free signals and the enhanced speech signals. Clearly, we can observe that the proposed algorithm (MLESS) has the best enhancement quality in all these tests. In the meantime,

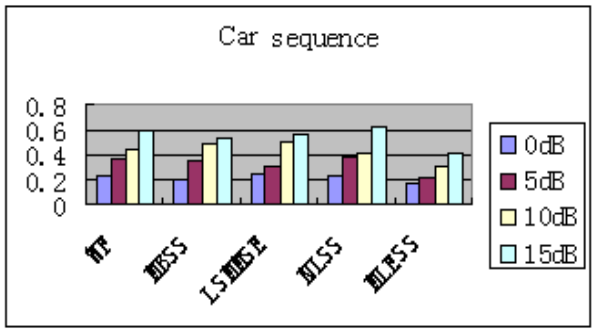

Figure 4. Performance comparison of different speech enhancement methods in the "car" sequence.

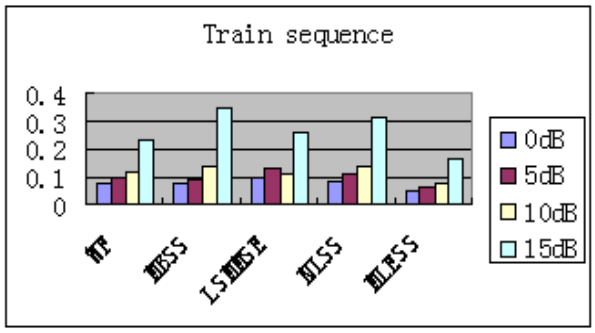

Figure 5. Performance comparison of different speech enhancement methods in the "train" sequence.

it also exhibits that other methods cannot hold consistent performance throughout the overall tests. In Figs. 4 and 5, we can find similar results to those of Fig. 3.

4. Conclusions and future work

We have described a technique for optimal speech enhancement in audiovisual indexing and retrieval, and the developed system has been evaluated in a number of experiments. We proposed a maximum likelihood estimate based spectral subtraction. This was undertaken using an iterative expectation-maximisation algorithm, allowing the probability distribution of useful signals to be approximated. We conducted experiments of speech enhancement in different noisy situations. Dynamic time warping based similarity check verified that the proposed algorithm had the best quality of speech enhancement, compared to other classical techniques. The future work is addressed on the applications of the proposed speech enhancement in audiovisual retrieval, and fusion of visual and auditory features in a single framework.

Acknowledgement

This work was supported by European Commission under Grant FP6-045189-STREP (RUSHES).

References

[1] Rushes project deliverable d5, requirement analysis and use-cases definition for professional content creators or providers and home-users. In *http://www.rushes-project.eu/upload/Deliverables/D5_WP1_ETB_v04.pdf*, August 2007.

[2] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of spech corrupted by acoustic noise. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 208–211, 1979.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, Series B 1977.

[4] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445, 1985.

[5] A. Hanbury and B. Marcotegui. Colour adjacency histograms for image matching. In *International Conference on Computer Analysis of Images and Patterns*, pages 424–431, 2007.

[6] Y. Hu and P. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Commun.*, 49(7-8):588–601, 2007.

[7] S. Kamath and P. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *International Conference on Acoustics, Speech, and Signal Processing*, page 4164, 2002.

[8] Y. Li, S. Narayanan, and C. Kuo. Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circuits Syst. Video Techn.*, 14(8):1073–1085, 2004.

[9] W. Qi, L. Gu, H. Jiang, X. Chen, and H. Zhang. Integrating visual, audio and text analysis for news video. In *International Conference on Image Processing*, pages III, 520–523, 2000.

[10] S. Tsekeridou and I. Pitas. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(4):522–535, 2001.

[11] H. Xu, A. Younis, and M. Kabuka. Automatic moving object extraction for content-based applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(6):796–812, June 2004.

[12] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng. Multi-sensory microphones for robust speech detection, enhancement and recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages III. 781–784, 2004.