Joint speech signal enhancement based on spectral subtraction and SVD filter

Wenkai Lu, Xuegong Zhang, Yanda Li*, Shen Liqin, Zhu Weibin**

*Institute of information processing, Department of automation, Key State Lab of Intelligent Technology and System, Tsinghua University, Beijing 100084, P.R. China.
Email: lwkmf@mail.tsinghua.edu.cn
** Speech Group, IBM China Research Lab.No 26, 6th Street, Shangdi, Haidian District Beijing, 100085, China

ABSTRACT

A joint speech signal enhancement based on singular value decomposition filter after spectral subtraction (SSVD) is proposed in this paper. The residual noise after spectral subtraction, which results for audible musical noise, is reduced further by SVD filter. The matrix size in spectral domain can be reduced half, and larger step-length adopted by SVD filter in spectral domain leads to lower cost, which make sure that the system can work in real-time. A novel speech/pause detector based on entropy (ESP) is proposed too. The new detector improves the performance of the whole noise suppression system significantly.

1. INTRODUCTION

The spectral subtraction offers the simple and computationally efficient tool for the suppression of an additive noise in a speech signal. The key idea of spectral subtraction is to estimate background noise and then to subtract this estimation from the noisy speech. The noise characteristics are usually updated during non-speech segments, the non-stationary noise will lead to serious residual noise. Even if the noise is stationary, the residual noise is unbearable in real applications.

Singular Value Decomposition has been proven to be an efficient tool for signal processing techniques: image coding, image filtering and signal enhancement, etc. SVD filter is used to suppress the residual noise after spectral subtraction here.

An accurate estimate of the noise is indispensable to obtain significant noise reduction. Many one-channel speech signal enhancement techniques estimate the noise characteristics in the speech pauses. Speech/pause detectors are the limiting parts of system for the suppression of additive noises in speech, because the quality of the detector determines the performance of the whole noise suppression system. A fast, reliable and robust speech/pause detector based on entropy is presented in this paper.

2. ENTROPY BASED SPEECH/PAUSE DETECTOR

It is well known that the amplitude spectrum of the speech signal is consisted by slight spikes with a few larger spikes almost. An entropy measure called varimax norm (Wiggins, 1978) can be used for speech signal detection in spectrum domain. The higher the varimax norm of a signal, the fewer spikes the signal has. Let the input signal is \( s(i) \), the varimax norm is defined as:

\[
V = \frac{\sum_{i} s(i)^4}{(\sum_{i} s(i)^2)^2}
\]

It is easy to see that only a little additive work is needed by ESPD. A few larger spikes mainly decide Varimax norm. So varimax norm is not varied almost in a certain SNR range. This is why only a fixed threshold classifier is needed in ESPD as long as same length of the segment is used. A female speech signal in chinese is used in our experiment. Normal random noise with different variance is added to form different SNR artificial noisy speech signal. Criteria (Petr Pollak, 1995) are used to evaluate the efficiency of ESPD.

- Correct speech detection rate-\( P(A/S) \)
- Correct pause detection rate-\( P(A/N) \)
- Correct detection rate-\( P(A) \)
  \[ P(A) = P(A/S) \times P(S) + P(A/N) \times P(N) \]
- Speech/pause resolution factor-\( P(B) \)
  \[ P(B) = P(A/S) \times P(A/N) \]

where \( P(S) \) and \( P(N) \) are rates of speech and pauses in the processed signal.

SNR range of the test signals is from –20DB to 20DB. Figure 1 shows the SNR dependance of the mean value of the criteria. The global correct detection rate is up to 0.9, which is superior to the results proposed in (Petr Pollak, 1995), when the SNR is 0DB.

Many experiments confirmed ESPD reliability and its ability to detect speech in strong noise with high probability.

3. NOISE SUPPRESSION

If the assumption of the only additive noise presence in the corrupted speech is made, i.e.

\[
x(n) = s(n) + d(n)
\]

where \( s(n) \) is the clean speech signal, and \( d(n) \) is the noise.

The use of one channel spectral subtraction method for noise suppression seems to be suitable with respect to its robustness.
and simplicity. This method is based on the estimation of the magnitude spectrum of enhanced speech signal because of the perceptual aspects of the human audible system. Good estimation of this spectrum is limited by the effect of musical tones.

The spectral subtraction we used here can be described by the equation

\[ \hat{S}(k) = |X(k)| - |D(k)| \]

After spectral subtraction, the spectral estimation of the speech signal is composed by two components

\[ \hat{S}(k) = |S(k)| + R(K) \]

where \( R(k) \) is the component corresponds to the residual noise.

In order to reduce the residual noise, we must estimate the magnitude of speech form the noisy signal \( \hat{S}(k) \). The noisy signal is treated as a vector in a \( T \)-dimensional space, with noise and speech components lying in orthogonal subspace. SVD filter is used to divide signal and noise subspace. From the noisy signal, the input matrix of SVD filter is constructed as

\[
A = \begin{bmatrix}
\hat{S}(I + 1) & \hat{S}(I + m + 1) & \ldots & \hat{S}(I + m \times N + 1) \\
\hat{S}(I + 2) & \hat{S}(I + m + 2) & \ldots & \hat{S}(I + m \times N + 2) \\
\ldots & \ldots & \ldots & \ldots \\
\hat{S}(I + T) & \hat{S}(I + m + T) & \ldots & \hat{S}(I + m \times N + T)
\end{bmatrix}
\]

with dimensions \( T \times N + 1 \), \( m \) is the overlapping step-length, \( I \) is the index of the frame, the total frame number is \( N+1 \), and \( T \geq N + 1 \).

The SVD of the matrix \( A \) is

\[ A = U \sum_{p} V_{p}^{T} \]

The largest singular components capture almost only signal information whereas the smallest ones contain almost only noise. A least squares/rank \( p \) estimate of matrix \( A \) is obtained by setting the \( N+1-p \) smallest eigenvalues to zero.

\[ A_{LS,p} = U \sum_{p} \left[ \begin{array}{c}
\sum_{p} V_{p}^{T} \\
0 \\
0
\end{array} \right]_{T \times (N+1-p)} \]

\[ = U \sum_{p} V_{p}^{T} \]

with \( \sum_{p} \) containing the largest singular values, and \( A_{LS,p} \) is the best rank-\( p \) approximation of matrix \( A \).

It is possible that there are some negative elements in \( A_{LS,p} \). To estimate the magnitude matrix \( B \) of speech from \( A_{LS,p} \), a transformation is needed to provide that the estimation is positive

\[ B(i, j) = \begin{cases} 
A_{LS,p}(i, j) & \text{if } A_{LS,p}(i, j) > \varepsilon \\
\varepsilon & \text{if } A_{LS,p}(i, j) \leq \varepsilon 
\end{cases} \]

with a small positive value \( \varepsilon \) as the threshold.

4. EXPERIMENT EXAMPLE

A female speech signal in Chinese is used in our experiment. The signals are sampled by frequency of 8 kHz and quantized to 16 bits. Normal random noise with different variance is added to form different SNR artificial noisy speech signal. The SNR different after and before signal enhancement (DSNR) is used to evaluate the proposed method.

\[ DSNR = SNR_{out} - SNR_{in} \]

The SNR is computed as

\[ SNR = 10 \log_{10} \frac{P_{S}}{P_{N}} = 10 \log_{10} \frac{P_{S}}{P_{X} - P_{S}} \]

where \( P_{X} \), \( P_{S} \), \( P_{N} \) is the PSD of noisy signal, speech and noise respectively.

In our experiments, the range of SNR of the input signal is from –10DB to 10DB. Figure 2 shows the SNR dependance of the mean value of DSNR, and Figure 3 shows the SNR dependance of the variance of DSNR with pause. Figure 4 shows the SNR dependance of the mean value of DSNR, and Figure 5 shows the SNR dependance of the variance of DSNR without pause.

The results in waveform are shown in Figure 6-8, in the case that the SNR of the input signal is –10DB, 0DB and 10DB. From top to bottom, there are the input speech, noisy speech, filtering signal and the residual signal between filtering signal and noisy signal.

5. CONCLUSION

In this paper, we presented a new technology for additive noise suppression in speech processing. The new type of speech enhancement algorithm can suppress the noise dramatically with lower cost. It can attenuate the musical noise caused residual noise after spectral subtraction.
The future works will include adaptive decision of $\text{Rank P}$, and the applications of the proposed method for robust speech recognition system.

6. REFERENCES


Figure 1: SNR dependence of the mean value of the criteria of ESPD

Figure 2: SNR dependence of the mean value of DSNR with pause

Figure 3: SNR dependence of the variance of DSNR with pause

Figure 4: SNR dependence of the mean value of DSNR without pause

Figure 5: SNR dependence of the variance of DSNR without pause
Figure 6: The results in waveform in the case that the SNR of the input signal is −10DB. From top to bottom, there are the input speech, noisy speech, filtering signal and the residual signal between filtering signal and noisy signal.

Figure 8: The results in waveform in the case that the SNR of the input signal is 0DB. From top to bottom, there are the input speech, noisy speech, filtering signal and the residual signal between filtering signal and noisy signal.

Figure 9: The results in waveform in the case that the SNR of the input signal is 10DB. From top to bottom, there are the input speech, noisy speech, filtering signal and the residual signal between filtering signal and noisy signal.