音声言語情報処理 29－11
（1999. 12. 20）

あらまし

キーワード

A nonlinear cepstral compensation method for noisy speech processing

Zhibin Pan[1)] Koji Kotani[1)] Tadahiro Ohmi[2)]

[1)] Department of electronic engineering, graduate school of engineering, Tohoku university

[2)] New industry creation hatchery center, Tohoku university

Aza-aoba 05, Aramaki, aoba-ku, Sendai, 980-8579, Japan

Tel: 022-217-7124, Fax: 022-217-9396, E-mail: pzb@sse.ecei.tohoku.ac.jp

Abstract

A nonlinear method is investigated in this paper to eliminate the influence of addictive noise on clean speech signal. The noise compensation is implemented through nonlinear averaging with exponential weights in time domain so that the most recent sampled datum is emphasized highest and previous ones exponentially less. To start this nonlinear algorithm, the first several hundred sampled data are pre-processed using conventional arithmetic averaging. In order to assess the effectiveness of this method, averaged distance of LPC cepstrum vector between noisy speech or exponentially averaged noisy speech and clean one is adopted as distortion measure. For 5 Japanese vowels, the cepstrum distortion of contaminated speech is obviously improved with S/N ratio ranging from 0 dB ~ 20 dB. For a population of 20 persons and each person with 5 speeches at S/N level of 0 dB ~ 20 dB, success rate of speaker identification in full matching way can be increased by 4%.

key words nonlinear average, cepstrum compensation, noisy speech

A nonlinear cepstral compensation method for noisy speech processing

Zhibin Pan[1)] Koji Kotani[1)] Tadahiro Ohmi[2)]

1) Department of electronic engineering, graduate school of engineering, Tohoku University
2) New industry creation hatchery center, Tohoku University

Aza-aoba 05, Aramaki, Aoba-ku,Sendai, 980-8579, Japan
E-mail: pzb@sse.ecei.tohoku.ac.jp

Abstract

A nonlinear method is investigated in this paper to eliminate the influence of addictive noise on clean speech signal. The noise compensation is implemented through nonlinear averaging with exponential weights in time domain so that the most recent sampled datum is emphasized highest and previous ones exponentially less. To start this nonlinear algorithm, the first several hundred sampled data are pre-processed using conventional arithmetic averaging. In order to assess the effectiveness of this method, averaged distance of LPC cepstrum vector between noisy speech or exponentially averaged noisy speech and clean one is adopted as distortion measure. For 5 Japanese vowels, the cepstrum distortion of contaminated speech is obviously improved with S/N ratio ranging from 0 dB ~ 20 dB. For a population of 20 persons and each person with 5 speeches at S/N level of 0 dB ~ 20 dB, success rate of speaker identification in full matching way can be increased by 4%.

1. Introduction

Speech and speaker recognition has made a great progress in recent years. Under laboratory condition, both of them have reached a rather high success rate. But in real world, because there inevitably exits noise from background and distortion due to transmission channel over telephone network, the performance of recognition degrades greatly.

To overcome this type of performance degradation, many noise reduction and transmission distortion compensation methods have been developed. For the former, the most famous one is spectrum subtraction (SSB)[2], which means to subtract the noise spectrum from noisy speech spectrum to get the true speech spectrum. For the latter, the cepstrum mean subtraction (CMS)[3], is well-known that refers to subtract the estimated mean value of cepstrum over a long enough period of time to compensate the transmission distortion.

In the case of noise reduction, SSB method has two problems that are not solved thoroughly. One is how to estimate the noise spectrum from noisy speech. Usually, the noise spectrum can be obtained by spotting the silent part of measured noisy speech and estimating its spectrum. Because the boundary between silent parts and unvoiced parts can not be divided clearly, the precision of estimated noise spectrum depends on how well the silent parts can be extracted. The other is that when the result of spectrum subtraction is too small, LPC filter will become unstable. This method is time-consuming.

In time domain, average is a common method to reduce noise. For example, N-term moving average can make noise reduced by $\sqrt{N}$. Because moving average gives the same weights to continuous neighboring N-terms, it also distorts the true speech signal obviously. Considering of strong correlation of speech signal, if not continuous neighboring N-terms but all terms previous are used and heavier weights are given to recent terms, it would be more reasonable. In this paper, average with exponential weights is investigated and the weights are adapted to the change of noisy speech signal.

Experiment results show that for 5 Japanese vowels "a, e, i, o, u" with S/N of 0 dB ~ 20 dB, exponential average can give distinct improvement for cepstrum. For a population of 20 persons and each person with 5 speeches that are taken in laboratory conditions and added white noise at S/N level of 0 dB ~ 20 dB, success rate of speaker identification in full matching way can be increased by 4% after exponential averaging.

2. Method of exponential average

In the case of addictive noise, noisy speech measured can be modeled as

$$x(n) = s(n) + w(n) \tag{1}$$

where x(n) is noisy speech signal corrupted by noise, s(n) is clean speech signal and w(n) is addictive white noise. n is time index of signal.

Noisy speech signal changes rapidly in time domain. To follow this change, while doing average, it would be better to give more weights to recent samples and less to the samples in past. The weights could be an exponential function with the most recent sample weighted the heaviest and previous samples exponentially less.

This algorithm can be implemented in one-step ahead average way recursively as shown in (2)

$$\begin{aligned} \hat{x}(n) &= sum(n) \\ sum(n) &= kx(n) + (1-k)sum(n-1) \end{aligned} \tag{2}$$

where, k is smoothing factor (0<k<1). The averaged output is computed through summing the current input sample multiplied by smoothing factor k and the previous summation multiplied by 1−k. Its calculation complexity is rather low.

By designating a initial value of previous summation to start the algorithm, it can continue to run till all samples are averaged. In this way, weights for the previous input samples decrease exponentially as new input sample is taken in. Because the weights vary with time exponentially and the current input sample is emphasized most, this algorithm has the ability to follow changes in noisy speech signal so that the degradation to true speech signal is less.

To carry out this algorithm, initial value of the previous summation and smoothing factor k has to be determined. Although the first sample can be used as initial value to start the algorithm, it will make the first sample be weighted too much. In this work, the linear weighting arithmetic average of first N samples is used as initial value of previous summation.

$$sum(0) = \frac{1}{N}\sum_{n=1}^{N} x(n) \tag{3}$$

Smoothing factor k determines to what extent past samples influence current averaged sample. And a small smoothing factor k results in a slow response to changes in noisy speech signal and large one allows rapid response to that. The better way to choose smoothing factor k is to make it adapt to the changes in the underlying noisy speech signal. If this algorithm is stable, one-step ahead errors will form a $N(0, \sigma^2)$ distribution and almost 95% errors should fall into $(-2\sigma, 2\sigma)$. However, if one-step ahead averaged value can not trace the change in noisy speech signal, large error will occur and exceed $(-2\sigma, 2\sigma)$. In this case, smoothing factor k should be adjusted to be able to follow this change.

When a new sample is available, the current smoothing factor k(n) as ratio of smoothed value of all errors SE(n) and smoothed value of all absolute errors $\Delta(n)$ given in (4) can be used to monitor the stability of averaging process.

$$\begin{aligned} k(n) &= \left|\frac{SE(n)}{\Delta(n)}\right| \\ SE(n) &= \gamma e(n) + (1-\gamma)SE(n-1) \\ \Delta(n) &= \gamma|e(n)| + (1-\gamma)\Delta(n-1) \end{aligned} \tag{4}$$

where SE(0)=0, $\Delta(0)=0$, γ is a smoothing constant, and e(n) is one-step ahead error.

The value of this monitor signal should be in between 0.2 ~ 0.5[5]. If it exceeds this range, it will be forced to fall into this range again by letting upper limit as 0.5 and lower limit as 0.2.

3. Experiments

For noisy speech signal, exponential average is conducted to show how much the cepstrum improvement can be reached. The noisy speech signal is generated by adding white noise to the clean one with S/N ratio ranging from 0 dB to 20 dB.

Cepstrum distance between clean speech and noisy speech or exponentially averaged noisy speech is used as a measure of distortion. If cepstrum distance becomes smaller after exponential

averaging, it means that exponential average can improve the distortion of cepstrum.

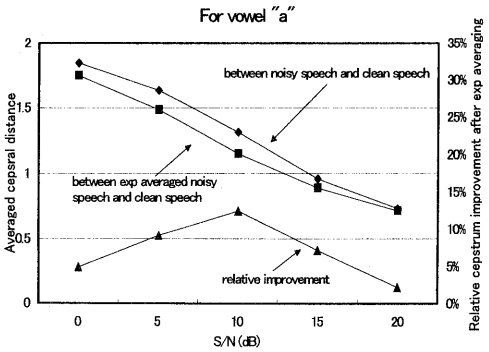
To assess the effect of exponential averaging, firstly, clean speech signal is divided into a series of frames and then for each frame LPC cepstrum vector is calculated using auto-correlation method. The analysis conditions for LPC cepstrum is summarized in Table1

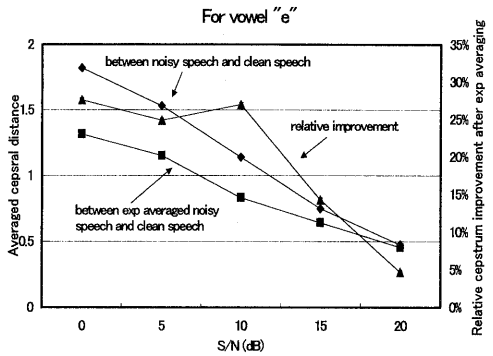
sampling rate	11.025 kHz
sampling precision	8 bit
low pass filtering	3 kHz
pre-emphasis	0.95
length of frame	30 ms
shift of frame	10 ms
window function	Hamming
LPC order	16
cepstrum number	21

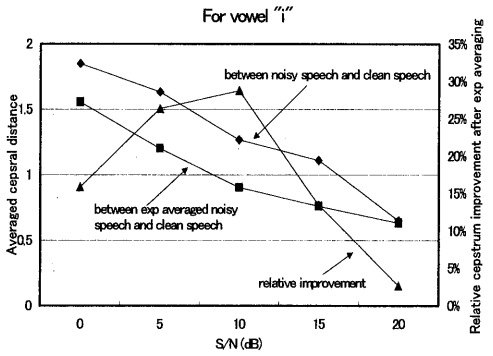
Table1 Analysis conditions for LPC cepstrum

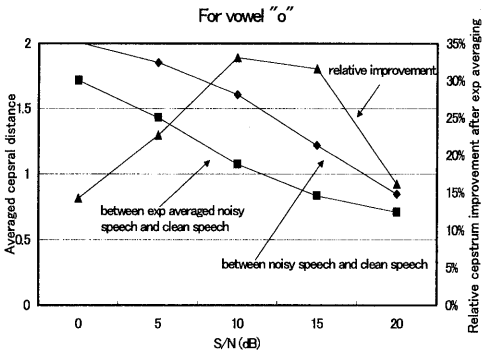
The LPC cepstrum vector of clean speech is used as baseline for comparison with that of noisy speech and exponentially averaged noisy speech[4]. Then, noisy speech signal is analysis in the same way and the distance between LPC cepstrum vector of noisy speech and that of clean one averaged over total speech is computed. This distance indicates how much noisy speech has been distorted relatively in cepstrum domain. Finally, noisy speech signal is exponentially averaged to compensate the effect of noise. The distance between LPC cepstrum vector of exponentially averaged noisy speech and that of clean one averaged over total speech is also calculated in the same manner as the case of noisy one. This distance measures the distortion of exponentially averaged noisy speech from clean speech as well.

Figure 1 shows the varying trend between absolute cepstrum distance, relative cepstrum distance improvement and S/N level for 5 Japanese vowels.









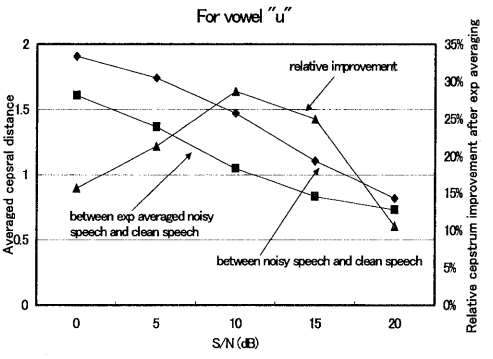


Figure1 Relation between absolute, relative cepstrum distance improvement and S/N ratio

From Figure1, it can be seen that cepstrum characteristics of all 5 Japanese vowels "a,e,i,o,u" are improved clearly. For absolute improvement, when S/N ratio becomes higher, the improved degree decreases. For relative improvement, with S/N ratio at medium level, the improved degree is better. Because noise is addictive and S/N ratio only changes noise level for a certain clean speech, degradation for clean speech is the same but for noise is greatly different at various S/N level due to averaging. Therefore, at higher noise level, the absolute improvement is better although cepstral distance itself is large. For relative improvement, because it is the ratio of absolute improvement and cepstral distance itself, the better one is at medium S/N level.

For noisy speeches with S/N of 0 dB ~ 20 dB, speaker identification among a population of 20 persons is also conducted. Everyone speaks 5 times. Each piece of speech is about 10 seconds and text-independent. Recorded 5 speeches are identified in full matching circulation way.

Classifier is realized through VQ and codebook is generated by Kohonen's LVQ. Assessment logic is decision by majority among mean, deviation of quantization error and S/N. The conditions for speaker identification using VQ are shown in Table2.

size of codebook	16
code dimension	20
initial codebook	random
times of learning	250
neighborhood set for updating	11
learning rate	linearly decreasing with leaning times
error measure	Euclidean distance

Table2 Conditions for speaker identification using VQ

Firstly, all noisy speeches are identified to get success rate. Depending on noise level, it is very different. Then, all noisy speeches are exponentially averaged and identified in the same way. The result is demonstrated in Figure2. It can be seen that success rate has been improved by at least 4% in the case of S/N is 20 dB.

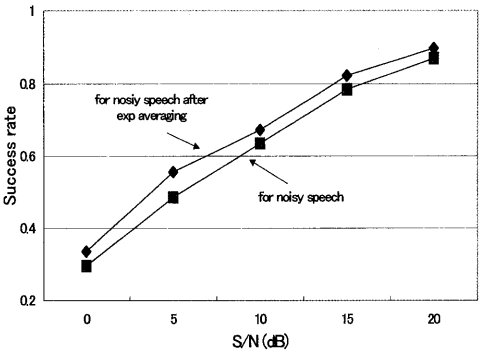


Figure2 Improvement of success rate for speaker identification after exp averaging

4. Conclusions

A average algorithm in time domain with exponential weights to compensate distortion in cepstrum domain due to addictive noise is investigated. This algorithm can follow the changes within noisy speech and it is computationally fast. Experimental results indicate that this method is effective to cepstrum improvement for 5 Japanese vowels "a, e, i, o, u" with S/N of 0 dB ~ 20 dB. And this algorithm is also applied to speaker identification among a small population to make success rate 4% up.

References

1) L.R.Rabiner and B.H.Juang, Fundamentals of speech recognition. Prentice-Hall, 1993.

2) S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoustics, Speech and Signal Processing, 27(2), 113-120,April 1979.

3) M.G.Rahim and B.H.Juang, "Signal bias removal for robust telephone based speech recognition in adverse environments," Proc. ICASSP, Adelaide, Australia, vol.1, pp.445-448, April 1994.

4) B.P.Milner and S.V.Vaseghi, "Comparison of some noise-compensation methods for speech recognition in adverse environments," IEE Proc. Vis. Image Signal Processing., vol.141, no.5, pp.280-288, October 1994.

5) B.Abraham and J.Ledolter, Statistical methods for forecasting. Wiley, 1983