NOISE-DRIVEN TEMPORAL TRAJECTORY FILTERING OF SPECTRAL PARAMETERS FOR ROBUST SPEECH RECOGNITION

Chandra Kant Raut Takuya Nishimoto Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
{raut, nishi, sagayama}@hil.t.u-tokyo.ac.jp

1. INTRODUCTION

Spectral parameter filtering such as RASTA [5], RASTA-like band-pass filtering and high-pass filtering operates on temporal dynamics of spectral parameters, and has been effective method to reduce channel distortions. Temporal derivatives (delta and delta-delta coefficients, that have proved as robust representation) and spectral mean normalization [4] are also equivalent to filtering that reject lower modulation frequencies from spectral parameters. Spectral normalization and parameter filtering assume that channel distortions are linear and additive noise is negligible [2]. As the additive noise and channel distortions are additive in different domains, they cannot be simultaneously suppressed by these methods [2].

In this paper, we extend the filtering technique of temporal trajectories to handle additive noise. The paper investigates the possibility of applying 'spectral subtraction' [3] filter to time-trajectories of spectral parameters and the problems that may arise.

2. SPECTRAL DYNAMICS

The spectral parameter $y(f, m\Delta T)$ of speech [Fig. 1] can be viewed as a time series that represents the temporal variations of the speech spectra. This time series will itself have spectral components, viz. modulation frequencies. The modulation frequencies can range up to half of analysis frame rate, e.g. 50 Hz for 10 ms of frame shift [2]. Houtgast *et al.* suggest that the relevant range of subband modulation frequencies for intelligible speech reproduction is approximately 0.4 to 20 Hz [2].

When noise trajectory's spectral component and speech trajectory's phonetically significant spectral component occupy different modulation frequencies, the effect of noise can be suppressed by linear filtering, using band-pass filters like RASTA and others. Furthermore, when noise trajectory's spectral component is concentrated at low frequency, high-pass filtering would be useful. However, it is difficult to deal with the case when noise trajectory's spectral component and speech trajectory's phonetically significant spectral component overlap significantly.

3. NOISE DATA-DRIVEN SPECTRAL PARAMETER FILTERING

The spectral energy of speech corrupted by additive noise, for f_i frequency bin in linear power spectral domain, is given by:

$$y_{f_i}(m\Delta T) = x_{f_i}(m\Delta T) + n_{f_i}(m\Delta T) \quad (1)$$

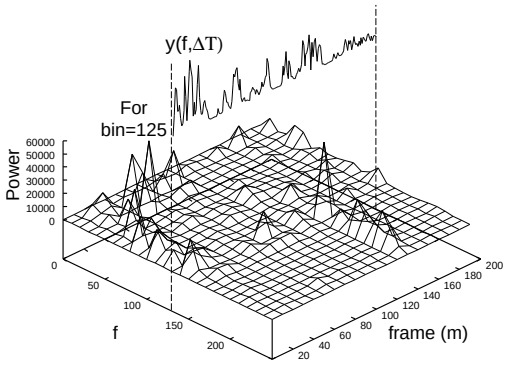

Fig. 1. Spectrogram and it's one frequency-bin

where m represents frame number and ΔT is frame shift. Parameters $y_{f_i}(m\Delta T)$, $x_{f_i}(m\Delta T)$ and $n_{f_i}(m\Delta T)$ are the temporal trajectories of noisy speech, clean speech and noise spectral energies respectively.

Taking Fourier transform of the trajectories gives:

$$Y_{f_i}(F) = X_{f_i}(F) + N_{f_i}(F) \quad (2)$$

where F represents the rate of change of spectral energies, modulation frequency.

Now, to filter out the modulation spectral component contributed by noise, we apply spectral subtraction filter to the trajectories:

$$|\widehat{X}_{f_i}(F)| = |Y_{f_i}(F)| - |N_{f_i}(F)| \quad (3)$$

$$|\widehat{X}_{f_i}(F)| = \left(1 - \frac{|N_{f_i}(F)|}{|Y_{f_i}(F)|}\right)|Y_{f_i}(F)| \quad (4)$$

Thus the frequency response of temporal trajectory filter is given by:

$$H_{f_i}(F) = \left(1 - \frac{|N_{f_i}(F)|}{|Y_{f_i}(F)|}\right) \quad (5)$$

In generalized form, (as in spectral subtraction):

$$H_b(F) = \left(Max\left(1 - \left(\frac{\beta(F)|N_b(F)|}{|Y_b(F)|}\right)^{\alpha}, \Theta\right)\right)^{\frac{1}{\alpha}} \quad (6)$$

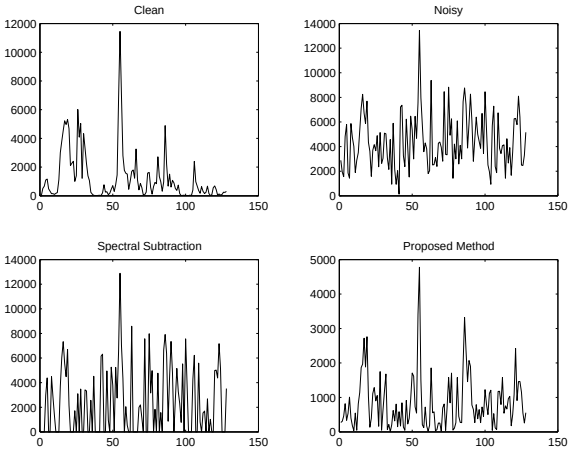

Fig. 2. Trajectories of single bin for clean speech and noisy speech; and trajectories estimated by spectral subtraction and noise-data driven temporal trajectory filtering.

where b represents a particular frequency bin in linear-spectral domain or mel-domain, β is scaling factor, Θ is threshold and $\alpha = 1$ for magnitude subtraction. Notice that, $\beta(F)$ is dependent on the modulation frequency.

The phase of spectral component of estimated trajectory for clean speech is taken same as that of corrupted trajectory:

$$\widehat{X}_{f_i}(F) = |\widehat{X}_{f_i}(F)|e^{j\phi_Y(F)} \tag{7}$$

The estimated trajectory is thus given by:

$$\hat{x}_{f_i}(m\Delta T) = \mathbf{F}^{-1}\left(\widehat{X}_{f_i}(F)\right) \tag{8}$$

where $\mathbf{F}^{-1}$ is inverse Fourier transform.

Figure 2 shows one-bin of the linear-spectrum of noisy speech and clean speech; and that estimated by spectral subtraction and noise-driven temporal trajectory filtering. As seen from the figure, the trajectory of spectrum-bin as estimated by proposed method is closer in nature to that of clean speech, however some good form of scaling or normalization will be required.

4. RESULTS

The performance of noise-driven temporal trajectory filtering (ND-TTF) was investigated on speaker dependent isolated word recognition task. The system was trained with 2620 words of same speaker from ATR A-Set speech database, and testing set comprised of 655 words of same speaker taken exclusively from the database.

The baseline system had 41 context independent CDHMMs with single mixture 123 total states; and 24-dimensional vector composed of MFCC and delta coefficients (with CMN applied) was used as input features to the recognizer. The feature extraction was done with the Hamming window of 25 ms size and frame shift of 10 ms (sampling frequency 16 kHz). The word accuracy for clean speech was 92.5 % with Julian 3.4 used as decoder.

Exhibition hall and station noise from JEITA database and whistle noise from RWCP database were added to test data at 10 dB and 20 dB SNR ; and recognition tests were performed directly and with noise-driven temporal trajectory filtering (ND-TTF).

Table 1. Word Accuracy (%)

Noise	dB	Direct	ND-TTF
Station	10	56.9	65.1
	20	82.0	85.2
Exhibition	10	57.6	58.7
	20	75.9	75.1
Whistle	10	45.8	59.4
	20	72.7	84.2

The result, as listed in Table 1, shows that the algorithm is performing better with the noise that has less overlapping modulation frequency than the noise that has modulation frequency significantly overlapping with speech (exhibition hall noise).

5. SUMMARY AND CONCLUSION

Noise-driven temporal trajectory filtering was found to be effective to suppress additive noise present in speech thus improving speech recognition performance. However, problems like run-time implementation, effective way of noise-sampling and updating etc. need to be addressed. Besides frequency bins in linear spectral domain are correlated to each other, and thus processing them independently may result in distortions. Some measures to reduce such the distortions need to be implemented. Furthermore, scaling of the estimated trajectory depending upon the SNR of the subband, and good choice of $\beta(F)$ are required for the algorithm to work well. The value of $\beta(F)$ should be chosen appropriately over the modulation frequency range, so that the significant range of modulation frequency of the spectral trajectory is well preserved.

Future work includes evaluation of algorithm in terms of spectral distortions as well as subjective test and its performance on other speech recognition tasks like AURORA-2J. Further improvement in algorithm and its implementation will be also considered.

6. REFERENCES

[1] C. Nadeu, B. H. Juang, "Filtering Spectral Parameters for Speech Recognition," in Proc. ICSLP94, pp. 1927-30, 1994.

[2] B. A. Hanson, T. H. Applebaum, and J. C. Junqua, "Spectral Dynamics for Speech Recognition under Adverse Conditions," in Automatic Speech and Speaker Recognition: Advanced Topics, *editors*: C. H. Lee, F. K. Soong and K. K. Paliwal, Kluwer Academic Publishers, 1996.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, Vol. 27, No. 2, pp. 113-120, 1979.

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the Speech Wave for Automatic Speaker Identification and Verification," in Journal of the Acoustical Society of America, 55(6), pp. 1304-1312, 1974.

[5] H. Hermansky and N. Morgan, "RASTA Processing of Speech," in IEEE Trans. on Speech and Audio Processing, 2(4), pp. 578-589, 1994.