RESEARCH NOTE

Noise Reduction Method Focusing on Spectral Envelopment and Fine Structure of Speech

Taichi Fukawa and Kenya Jin'no

Graduate School of Engineering, Tokyo City University, Tokyo 158-8557, Japan

E-mail: g2181479@tcu.ac.jp, kjinno@tcu.ac.jp

Abstract We propose a method to enhance noise reduction performance by separating a speech spectrum into spectral envelopes and fine structures using cepstrum analysis and linear predictive coding (LPC) analysis, and removing noise using an autoencoder (AE). A technique for removing noise from the spectrum of noise-containing speech is to use AE to reconstruct the spectrum of speech through the latent variables of the speech. We focused on spectral envelopes and fine structures that constitute speech, and improved the independence between latent variables in AE to reconstruct the speech spectrum by separating them in advance. In this way, we confirmed that the performance of noise reduction was improved in exchange for a slight decrease in the reproducibility of speech spectra when cepstrum analysis was used. It was also confirmed that cepstrum analysis was superior to LPC analysis in noise reduction.

Keywords: noise reduction, spectral envelope, fine structure, autoencoder, cepstrum analysis, LPC analysis

1. Introduction

Speech recognition systems improve analysis performance by extracting only the speech signal from a signal containing noise as a preliminary step. A spectrum of speech consists of a convolution of components with different properties, called spectral envelopes and fine structures [1], but conventional methods perform noise reduction with these components mixed [2]. Focusing on this convolutional property, we propose a noise reduction method that separates the spectrum into spectral envelopes and fine structures. After the separation, the spectrum is resynthesized with the speech spectrum by removing noise with an autoencoder (AE), which reproduces the spectrum via latent variables.

2. Methods

2.1 Input signal model

Since the input signal waveform has an additive relationship between the energy of the speech waveform w_{voice} and that of the noise waveform w_{noise} , the input signal waveform w to the noise reduction system is as follows:

$$w = w_{\text{voice}} + w_{\text{noise}} \quad (1)$$

The vocal tract from the vocal cords to the lips, which is the structure that produces speech, is modeled as a series of circular acoustic tubes of different diameters (Fig. 1) [2][3]. By inputting an impulse train in the case of a voiced sound or white noise in the case of a silent sound as the input signal $e(t)$ to the acoustic tube, the speech signal $w(t)$ is generated by the frequency filters of spectral envelopes and fine structures, respectively [1]. Spectral envelopes affect

Fig. 1 Structure of speech production

pronunciation, such as vowels and consonants, so when they change, the pitch remains the same but the pronunciation changes. In contrast, the fine structures affect the pitch of the sound, so when they are changed, the pitch changes while the pronunciation remains the same.

The noise waveform w_{noise} generated by unspecified environmental noise can be regarded as white Gaussian noise by the central limit theorem. White

Gaussian noise is a waveform whose amplitude follows a Gaussian distribution with 0 as the mean and whose frequency spectrum is of uniform power.

2.2 Separation into spectral envelopment and fine structure

Although the spectrum obtained by discrete Fourier transform (DFT) can be used to analyze stationary speech signals, it is difficult to analyze speech whose frequency characteristics change with time. Therefore, in order to analyze the frequency response of speech, which is nonstationary, the frequency spectrum is obtained by short-time Fourier transform (STFT) and only the amplitude component is extracted. After noise reduction, the complex spectrum is obtained by combining the amplitude and phase components, and the speech is resynthesized.

When the spectrum is plotted on a plane graph with frequency on the horizontal axis and logarithmic power on the vertical axis, the component that shows very fine changes is the fine structure, and the component that shows gradual changes, such as connecting the centers or peaks of the wave, is the spectral envelope. Since these boundaries are not clearly defined, assumptions are made to separate them.

2.2.1 Cepstrum analysis

In cepstrum analysis, the logarithmic spectrum plotted on the frequency–log power axes is considered as a waveform, and spectral envelopes are separated as low-frequency components and microstructures are separated as high-frequency components [4]. A waveform converted by DFT as a time waveform from a graph with log power on the vertical axis is called a cepstrum, derived from the word “spectrum”. The unit of the horizontal axis of the cepstrum is time, but since it is different from ordinary time, it is called quefrency, which is derived from the word “Frequency”, and the equivalent of a filter on the cepstrum is called lifter. Since spectral envelopes and fine structures have a multiplicative relationship in the amplitude spectrum, the use of a logarithmic spectrum makes the relationship additive, and each can be extracted by the lifter.

2.2.2 LPC analysis

To model the coupled acoustic tube in linear predictive coding (LPC) analysis, the autoregressive (AR) process in Eq. (2) is assumed to hold for the speech signal $w(t)$.

$$\sum_{i=0}^n \alpha_i w(t-i) = e(t) \quad (2)$$

The autocorrelation coefficient at $t = 0$ is $\alpha_0 = 1$. The linear prediction error $e(t)$ is the source signal of the

vocal cords, and the LPC coefficient α_i is determined to minimize $e(t)$.

Replacing Eq. (2) with the z-transform, we obtain the following:

$$\sum_{i=0}^n \alpha_i W(z) z^{-i} = E(z) \quad (3)$$

$$W(z) = \alpha_0 W(z) = \frac{1}{\sum_{i=1}^n \alpha_i W(z) z^{-i}} E(z) \quad (4)$$

$$W(z) = E(z) \prod_{i=1}^n \left(\frac{1}{1 + \beta_i z^{-1}} \right) \quad (5)$$

Equation (5) can be viewed as a series of n acoustic tubes, each with its own resonant properties.

If we assume that $H(z) = \sum_{i=1}^n \alpha_i W(z) z^{-i}$, then Eq. (4) can be regarded as a system given by the transfer function $1/H(z)$.

$$W(z) = \frac{1}{H(z)} E(z) \quad (6)$$

This gives us the frequency response of $1/H(z)$ as an amplitude spectrum.

α_i is obtained from the audio waveform frame w_n as a solution to a linear equation by the Levinson–Durbin recursion method [5]. The value of n assumed in the AR process corresponds to the order of the spectrum when viewed as a curve with a polynomial function, so if n is small, it is a low-order function. The frequency response curve on this spectrum is taken as a spectral envelope, and the fine structure can be obtained by dividing the spectral envelope from the power of the original spectrum.

Fig. 2 Extraction of spectral envelopes of speech

Both cepstrum analysis and LPC analysis are characterized by the fact that they can be obtained with a small amount of computation. As can be seen from Fig. 2, the spectral envelope obtained by cepstrum analysis is characterized by a curve that passes through the center of the wave of the logarithmic power spectrum. In LPC analysis, a curve that connects the power peaks of the spectrum is obtained to reflect the formant resonance.

2.3 Noise reduction

Spectral envelopes that change shape depending on pronunciation, such as vowels and consonants, may have latent variables that generate the pattern. Since the stripes of the microstructure show the same shape at a specific period, there may be a latent variable that specifies the period.

AE is used to learn nonlinear transformations using neural networks to reconstruct the input data via latent variables [6]. To learn a noise-resilient transformation, an amplitude spectrum generated from a signal waveform containing white Gaussian noise is given as the input, and network weights are learned using the noise-free speech-only spectrum as the correct data. Since white Gaussian noise is difficult to compress, it is expected that the encoder in Eq. (7) will be learned in such a way as to weaken the effect of the noise on the latent variable.

$$\text{AE} : \mathbf{s}_n \mapsto \mathbf{s}_{(\text{voice}),n} \quad (7)$$

By inputting a noisy spectrum into this network, the spectrum with the noise removed is obtained as an output as shown in Eq. (8).

$$\mathbf{s}'_n = \text{AE}(\mathbf{s}_n) \quad (8)$$

In our experiment, unlike the conventional method [6], which uses a single AE, we use separate AEs to learn denoising transformations for spectral envelopes and fine structures separated by cepstrum or LPC analysis. Since the speech data is separated using the properties of speech, the learning performance is expected to improve owing to the reduction in the number of speech features contained in each component.

The denoised spectral envelope and fine structure are combined by the inverse operation of the cepstrum or LPC analysis used for the separation, and the denoised spectrum is reconstructed.

In AE, the network from the input layer to the middle layer is called the encoder, and the network from the middle layer to the output layer is called the decoder. By reducing the number of dimensions in the middle layer to less than that of dimensions in the input layer, the encoder learns the network to find the latent variables required to restore the spectrum in the decoder, and the decoder learns the network to restore the spectrum from the latent variables.

Here, the encoder and decoder are symmetric with respect to the intermediate layer, and the structure from the input layer to the intermediate layer is represented as $[N_0, N_1, \dots, N_n]$.

3. Improvement to High Definition

AE can obtain values closer to the true latent variables of the learning target by reducing the number

of latent variables, but it may not be able to reproduce original spectrogram with sufficient quality owing to lack of information when compression is insufficient. Therefore, we use a high-definition segmentation method, U-Net (Fig. 3) [7], and provide a skipping structure in which the output at each layer of the encoder is input to each layer of the decoder to increase the resolution of the spectrum of the output.

Fig. 3 Network structure of U-Net

Since the input is passed to the output without the latent variable, there is a disadvantage that noise appears in the output spectrum, but it works effectively when the effect of noise in the input is small compared with the error generated by the reconstruction of the spectrum by the latent variable.

4. Results

To learn AE, we use the JVS corpus [8], which contains 100 phonemically balanced sentences of Japanese, recorded with a sampling frequency of 24,000 Hz and a bit depth of 16 bits, for 100 people. 64 utterances each from 64 people are used for training, 16 utterances each from 16 people are used for validation, and 16 utterances each from 16 people are used for evaluation, with no overlap in speakers and corpus sentences. White Gaussian noise is added to the speech waveform with a signal-to-noise ratio (SNR) of 4, which is a measure of the root mean square (RMS). 2048 samples per frame and 128 frame intervals are used for STFT to obtain spectrograms. Owing to the different lengths of the audio, we use the audio trimmed to 5.44 s. The layered structure of AE that received the highest rating in the preliminary experiments shown in Table 1 was used.

The evaluation is based on the mean square error (MSE) of the spectrogram, the signal-to-distortion ratio (SDR) of BSS_EVAL [9], and the SNR of the speech waveform. The evaluation scores of noise reduction by AE using the fully connected and convolutional layers are shown in Table 2.

Regardless of the type of AE layer or the separation method, U-Net scored higher than AE in most

Table 1 Layer structures used in AE for noise reduction

Layer Type	Separation Method	Component	Layer Structure
Fully connected	None (conventional method)	All	[128, 64, 32, 16]
	Cepstrum analysis	Spectral envelope Fine structure	[64, 32, 16, 8, 4] [256, 128, 64, 32, 16]
	LPC analysis	Spectral envelope Fine structure	[128, 64, 32, 16, 8] [640, 512, 384, 256, 128]
Convolutional	None (conventional method)	All	[128, 256, 512]
	Cepstrum analysis	Spectral envelope Fine structure	[8, 16, 32, 64] [64, 128, 256, 512]
	LPC analysis	Spectral envelope Fine structure	[32, 64, 128] [24, 48, 96]

Table 2 Evaluations of noise reduction performance

Layer Type	Separation Method	Model	MSE	SDR [dB]	SNR [dB]	s^2 (SNR)
Fully connected	None (conventional method)	AE	7.30	4.33	4.43	0.82
		U-Net	8.46	5.73	5.56	6.12
	Cepstrum analysis	AE	13.14	3.48	3.14	9.48
		U-Net	5.61	6.78	6.63	3.52
	LPC analysis	AE	14.25	2.49	2.52	1.57
		U-Net	7.61	5.07	5.12	3.25
Convolutional	None (conventional method)	AE	3.68	7.86	7.91	1.12
		U-Net	3.84	7.71	7.71	1.71
	Cepstrum analysis	AE	9.90	4.68	4.13	3.60
		U-Net	2.79	9.24	9.14	3.70
	LPC analysis	AE	6.29	5.26	5.35	1.20
		U-Net	3.88	7.28	7.39	1.20

cases. This means that the error generated by AE in reproducing the speech spectrum is larger than that generated by the noise input through the skip structure of U-Net.

We confirm that when using AE, the score is higher using the conventional method without separation, but when using U-Net, the noise removal performance can be significantly improved by using cepstrum analysis. This is due to the fact that distortions contained in the spectral envelope and fine structure, which are product relations, are amplified in the spectral resynthesis when the reproducibility by AE is insufficient, and the reason for this result is that the effect of distortions was sufficiently low when using U-Net, which has sufficient reproducibility.

It was also confirmed that regardless of the type of AE layer, the noise reduction performance was always higher when using cepstrum analysis than when using LPC analysis, and exceeded the score of the conventional method only when using cepstrum analysis.

5. Conclusions

We compared the separation method based on cepstral analysis with that based on LPC analysis, and

confirmed that the former method has better noise reduction performance. We also confirmed that the proposed method of separating spectral envelopes and fine structures by cepstrum analysis has the potential to improve the noise reduction performance of speech signals at the expense of a slight decrease in the reproducibility of the speech spectrum.

Acknowledgments

This work was supported by JSPS KAKENHI Grants-in-Aid for Scientific Research (C) Numbers 19K12163 and 20K11978. Part of this work was carried out under the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University.

References

- [1] R. D. Kent and C. Read: The Acoustic Analysis of Speech, Kaibundo, Translated by T. Arai and T. Sugahara, 1996.
- [2] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar and T. Weyde: Singing voice separation with deep

U-Net convolutional networks, 18th ISMIR, pp. 745–750, 2017.

- [3] M. M. Sondhi: Resonances of a bent vocal tract, *The Journal of the Acoustical Society of America*, Vol. 79, No. 4, pp. 1113–1116, 1986.
- [4] S. Itahashi: *Speech Technology*, Morikita, 2005.
- [5] MathWorks: Solve linear system of equations using Levinson-Durbin recursion, December 4, 2020, <https://www.mathworks.com/help/signal/ref/levinson.html>.
- [6] S.-W. Fu, Y. Tsao and X. Lu: SNR-aware convolutional neural network modeling for speech enhancement, *Proc. Interspeech*, pp. 3768–3772, 2016.
- [7] O. Ronneberger, P. Fischer and T. Brox: U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Cham, Springer, pp. 234–241, 2015.
- [8] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari: JVS corpus: free Japanese multi-speaker voice corpus, arXiv:1908.06248, 2019.
- [9] E. Vincent, R. Gribonval and C. Févotte: Performance measurement in blind audio source separation, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 4, pp. 1462–1469, 2006.

Taichi Fukawa received his B.E. degree in engineering from Tokyo City University, Japan in 2021. Currently, he is a graduate student of Tokyo City University. His research interests include machine learning, signal processing, image processing and speech processing.

Kenya Jin'no received his B.E., M.E. and Ph.D. degrees in electrical engineering from Hosei University, Tokyo, Japan, in 1991, 1993 and 1996, respectively. In 2018, he joined the Department of Intelligent Systems, Tokyo City University, Tokyo, Japan where he is currently a professor. His research interests include machine learning, soft computing, high-dimensional nonlinear systems, image processing, complex systems and evolutionary computation. Dr. Jin'no is a member of RISP, IEICE (Fellow), IEEE, INNS, SICE, JSEC and IPSJ.

(Received May 11, 2021; revised July 26, 2021)