Vol.2010-SLP-82 No.5
2010/7/23

頑健な音声認識のためのウエーブレットパラメータの最適化に基づく残響抑圧

ゴメス・ランディ †1　　河原　達也†1

本稿では、音声認識のためのウエーブレットに基づく残響抑圧法について述べる。残響抑圧は、遅い残響成分の影響を抑圧するように、ウエーブレット係数をウイナーゲインでフィルタリングすることで行なう。音響モデルの尤度に基づいてウエーブレットのパラメータを最適化することで、元音声と遅い残響成分をそれぞれ効果的に推定することができ、これにより、音声認識に適した残響抑圧のためのウイナーゲインを求めることができる。このウイナーゲイン自体も音響モデルの尤度を用いて調整することで、さらに残響抑圧が改善される。残響のある実データを用いた評価実験により、有意な音声認識精度の改善を得ることができた。

Robust Speech Recognition using Optimized Wavelet-based Dereverberation

Randy Gomez †1 and Tatsuya Kawahara†1

This paper presents an improved wavelet-based dereverberation method for automatic speech recognition (ASR). Dereverberation is based on filtering reverberant wavelet coefficients with the Wiener gains to suppress the effect of the late reflections. Optimization of the wavelet parameters using acoustic model enables the system to estimate the clean speech and late reflections effectively. This results to a better estimate of the Wiener gains for dereverberation in the ASR application. Additional tuning of the parameters of the Wiener gain in relation with the acoustic model further improves the dereverberation process for ASR. In the experiment with real reverberant data, we have achieved a significant improvement in ASR accuracy.

†1 京都大学 学術情報メディアセンター
ACCMS, Kyoto University

1. Introduction

Acoustic degradation of the speech signal caused by reverberation poses a problem in distant-talking speech recognition applications. The observed signal in the microphone is smeared with both the effects of early and late reflections. We have proposed a dereverberation approach[1)2)] that suppresses the late reflection of the reverberant signal by means of multi-band spectral subtraction. This method is analogous to the multi-band spectral subtraction steered by multi-step linear prediction[3)]. In[1)2)], the power estimate of the late reflection is crucial in the dereverberation process. However, there is no straightforward means of accurately estimating it, as its characteristics vary accordingly as a function of the room characteristics and the energy of the preceding speech-frame segments.

In this paper, we propose a wavelet-based dereverberation approach optimized for ASR as shown in Fig. 1. First, we estimate the room reverberation time T_{60} to obtain the room impulse response (RIR). Then, we reproduce the reverberant data set and optimize separate wavelet parameters (i.e. scale and shift) for speech and late reflection, respectively. The optimization process is based on improving the model likelihood of the speech recognizer through offline training. In the actual dereverberation process, wavelet filtering is employed by weighting the reverberant wavelet coefficients with multi-band Wiener gains. In calculating the Wiener gains, we estimate the clean speech and the late reflection power using the optimized wavelet parameters. Then, we tune the parameter of the Wiener gain based on the acoustic model likelihood. During testing, the optimized wavelet parameters and the tuned parameter of the Wiener gain

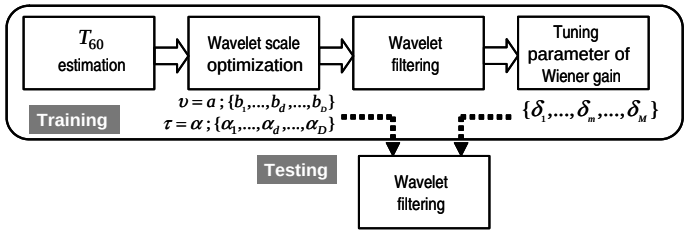

図 1 Block diagram of the proposed method.

are used for dereverberation through wavelet filtering.

The paper is organized as follows; Section 2 gives an overview of the reverberant model and the concept of the dereverberation approach we adopt. Section 3 presents the proposed method of wavelet-based dereverberation. Experimental conditions and results are given in Section 4, and we will conclude this paper in Section 5.

2. Reverberant Speech Model

2.1 Early and Late Reflection

The spectrum of the reverberant signal (frequency f, time t) is given as,

$$X(f) \approx S(f)H(f) \tag{1}$$

where $X(f)$, $S(f)$ and $H(f)$ are the frequency components of the reverberant signal, clean speech signal and the room impulse response (RIR), respectively. The reverberation effect can be decomposed into early and late reflections. The early reflection is due to the direct signal and some reflections that occur at earlier time and can be treated as short-period noise. The late reflection, whose effect spans over frames can be treated as long-period noise. The RIR h can be expressed with early h_E and late h_L components as follows,

$$h_E(t) = \begin{cases} h(t) & t < T \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$h_L(t) = \begin{cases} h(t+T) & t \geq T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where T denotes the frame length. Eq. (2) and (3) characterize both the short and long-period effects of the reverberant signal. The short term fourier transform (STFT) of the reverberant signal can be expressed in terms of early and late reflections as,

$$\begin{aligned} X(f,t) &= S(f,t)H(f,0) + \textstyle\sum_{d=1}^{D-1} S(f,t-d)H(f,d) \\ &= \quad X_E(f,t) \quad + \quad X_L(f,t) \end{aligned} \tag{4}$$

where $H(f,0)$ is the RIR in-frame effect to the speech signal $S(f,t)$ due to $h_E(t)$. We denote this as early reflection $X_E(f,t)$. The second term $X_L(f,t)$ referred to as the late reflection can be viewed as smearing of the clean speech by $H(f,d)$ which corresponds to the d frame-shift effect of the RIR due to $h_L(t)$. D is the number of frames over which the reverberation (smearing) has an effect and is related with the reverberation time T_{60}. The early reflection is mostly addressed through Cepstral Mean Normalization (CMN) in the ASR system as it falls within the frame. Thus, we focus on suppressing only the effect of the late reflection.

2.2 T_{60} Estimation

The HMM representation of a speech signal is of low resolution compared to the actual RIR. Thus, in HMM-based ASR applications, it may be sufficient to use T_{60} estimate in describing the RIR characteristics of a room[5]. The multiple reflections of sound can be described by a decaying acoustical energy given as

$$h^2(l) \approx e^{(\ \ln\ (10)/T_{60})\ l}, \tag{5}$$

where l is the discrete time sample, and T_{60} is the reverberation time. Fig. 2 illustrates the process of T_{60} estimation. First, we generate reverberant data x^{T60_1} ... x^{T60_K} based on Eq. (5) and train GMM with 64 mixtures for each: μ_{rev_1} ... μ_{rev_K}. In the actual T_{60} estimation, the likelihood scores are evaluated against μ_{rev}, and the subsequent T_{60} that results in the highest likelihood score is selected. Although this can be more accurately measured through physical measurement[8], it may be impractical and inconvenient whenever the room characteristics change. By using the T_{60} estimate, we can synthetically generate the RIR using Eq. (5). With the RIR, h_L is identified experimentally in our previous work[1)4)].

3. Wavelet Filtering for Dereverberation

3.1 Wavelet Parameter Optimization

The advantage of wavelet over the short-time fourier transform (STFT) is its flexibility to analyze the spectral component and detect changes across the spectrum[6]. A wavelet is generally expressed as

$$\Psi(\upsilon,\tau,t) = \frac{1}{\sqrt{\upsilon}}\Psi\left(\frac{t-\tau}{\upsilon}\right), \tag{6}$$

where t denotes time, υ and τ are the scaling and shifting parameters respectively.

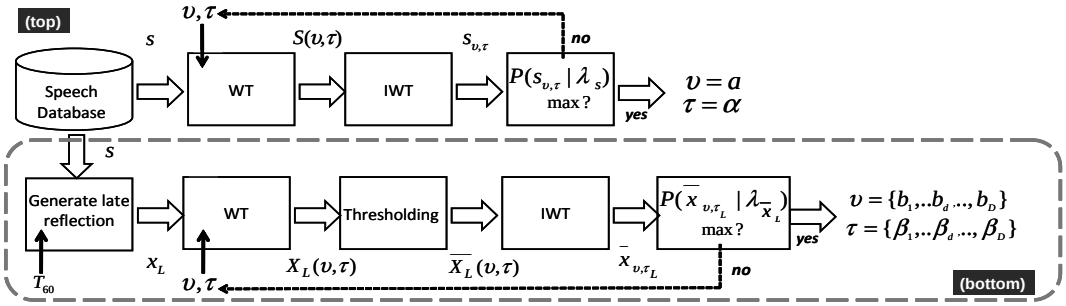

図 3 Wavelet optimization scheme.

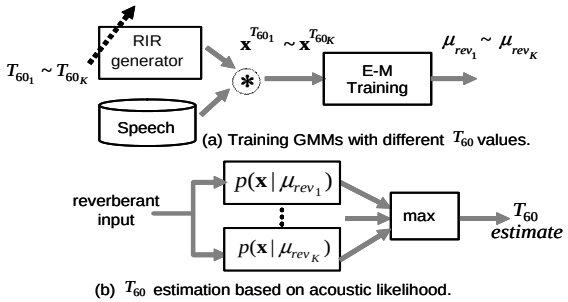

図 2 Room impulse response approximation

$\Psi\left(\frac{t-\tau}{\upsilon}\right)$ is often referred to as the mother wavelet. Assuming that we deal with real-valued signal, the wavelet transform (WT) is defined as

$$F(\upsilon,\tau) = \int f(t)\Psi(\upsilon,\tau,t)dt, \tag{7}$$

where $F(\upsilon,\tau)$ is the wavelet coefficients and $f(t)$ is the time-domain function. Unlike the constant window analysis in STFT, WT offers the flexibility of shifting and scaling the mother wavelet shown in Eq. (6). Shifting the wavelet may delay or hasten its offset. The scale parameter controls the degree of representation of the feature parameters of the signal of interest. Thus, with an appropriate training algorithm we can optimize τ and υ so that the wavelet captures specific characteristics of a certain signal of interest. The resulting wavelet is sensitive in detecting the presence of this signal given any arbitrary signal. In our proposed dereverberation approach, we are interested in detecting the power of clean speech and late reflection given a reverberant signal.

We optimize the wavelet to detect clean speech and late reflection separately based on the acoustic model likelihood as shown in Fig. 3. In ASR, we assume that the speech does not vary for a certain time-frame. Thus, optimizing a single wavelet template for speech will be sufficient. In Fig. 3 (top) we illustrate the optimization of the wavelet for clean speech. Wavelet coefficients $S(\upsilon,\tau)$, extracted through Eq. (7), are converted back to time domain $s_{\upsilon,\tau}$. Likelihood scores are computed using the clean speech acoustic model λ_s. The process is iterated, adjusting υ and τ. The corresponding υ=a and τ=α that result to the highest score are selected. In the case of the late reflection in Fig. 3 (bottom), D templates are to be optimized for both scale $(\upsilon_1,...\upsilon_D)$ and shift $(\tau_1,...,\tau_D)$. These correspond to D frames that cause smearing as depicted in Eq. (4). We note that the effect of smearing is not constant, thus D templates are created. As discussed in Section 2.2, we can avail of the late reflection coefficients h_L from Eq. (5) after estimating T_{60}[1)4)]. Then, late reflection observations x_L are generated by convolving the clean speech with h_L. Next, wavelet coefficients $X_L(\upsilon,\tau)$ are extracted through WT (Eq. (7)). To make sure that $X_L(\upsilon,\tau)$ is void of speech characteristics, thresholding is applied to $X_L(\upsilon,\tau)$. Speech energy is characterized with high coefficient values[9)10)] and thresholding sets these coefficients to zero,

$$\bar{X}_L = \begin{cases} 0 & ,\mid X_L \mid > \ thresh \\ X_L & ,\mid X_L \mid < \ thresh \end{cases} \tag{8}$$

where *thresh* is calculated similar to that of[9] using

$$thresh = \sigma\sqrt{2\ \log(L)}, \tag{9}$$

where L is the length of the late reflection with variance σ^2 over the span of D. The thresholded signal is converted back to time domain $\bar{x}_{\upsilon,\tau_L}$ and evaluated against a thresholded late reflection model $\lambda_{\bar{x}_L}$. The parameters υ and τ are adjusted and the corresponding υ=$\{b_1,...b_D\}$ and τ=$\{\beta_1,...\beta_D\}$ that results to the highest likelihood score is selected. We note that the acoustic model λ_s is trained with clean speech data, while $\lambda_{\bar{x}_L}$ uses the synthetically generated late reflection data with thresholding applied.

3.2 Wavelet Filtering

We have expanded the multi-band wavelet domain filtering[7] to address the dereverberation problem[12]. The general expression of the Wiener gain at band m[12] is expressed as

$$\kappa_m = \frac{S(\upsilon,\tau)_m^2}{S(\upsilon,\tau)_m^2 + \delta_m X_L(\upsilon,\tau)_m^2}, \tag{10}$$

where $S(\upsilon,\tau)_m^2$ and $X_L(\upsilon,\tau)_m^2$ are wavelet power estimates for the clean speech and the late reflection, respectively. By using the optimized values for υ and τ discussed in Section 3.1, we can estimate these parameters directly from observed reverberant signal $X(\upsilon,\tau)$. Thus, the speech power estimate becomes

$$S(\upsilon,\tau)_m^2 \approx X(a,\alpha)_m^2, \tag{11}$$

and the late reflection power $X_L(\upsilon,\tau)_m^2$ estimate

$$X_L(b_d,\beta_d)_m^2 \approx \begin{cases} X(b_1,\beta_1)^2\ ,\ d=\ 1 \\ \dfrac{\sum_{k=1}^{d-1} X(b_k,\beta_k)^2}{d-1} + X(b_d,\beta_d)_m^2\ , \\ \quad \text{otherwise} \end{cases} \tag{12}$$

where d is the d-th frame template (for k:1,...,D). We note that the contribution of the preceding frames is also considered in Eq. (12). If the late reflection power estimate is greater than the estimate of the speech power, then κ_m for that band may be set to zero or a small value. Due to the non-stationary characteristics of the late reflection, a tuning parameter δ_m is introduced to compensate the estimation error of $X_L(\upsilon,\tau)_m^2$. Wavelet filtering is carried out by weighting the reverberant wavelet coefficients with the Wiener gains as,

Methods	**200 msec**	**600 msec**
(A) No processing; clean model	68.6 %	21.4 %
(B) No processing; reverb model	75.4 %	32.1 %
(C) Improved thresholding[10]	77.3 %	50.6 %
(D) Improved thresholding[10] + wavelet optimization	79.1 %	54.0 %
(E) Extrema clustering[11]	78.4 %	59.7 %
(F) Extrema clustering[11]+ wavelet optimization	80.8 %	62.9 %
(G) Wavelet Filtering	81.5 %	64.5 %
(H) Wavelet Filtering + wavelet optimization	83.2 %	68.6 %

表 1 ASR results in Word Accuracy

$$X(\upsilon,\tau)(enhanced) = X(\upsilon,\tau)_m\ \ .\ \ \kappa_m. \tag{13}$$

The Wiener weighting κ_m dictates the degree of suppression of the late reflection to the observed signal. We note that the optimized υ and τ are only used in calculating the Wiener gains. The enhanced wavelet coefficients are converted back to the time domain through IWT. In our previous work[12], the wavelet parameters are not optimized to track the clean speech and the late reflection given a reverberant observation. The method[12] relies solely in tuning of δ_m to compensate the estimation error, which is reviewed in the next subsection.

3.3 Tuning Parameter of Wiener Gain

We also introduce a multi-band parameter δ_m (for band m: 1,...,M) to tune the Wiener gain in Eq. (10). These values are adjusted and selected in relation to the acoustic model likelihood. Thus, a set $\{\delta_1,...,\delta_m,...,\delta_M\}_{opt}$ is optimized through maximum likelihood criterion as described in[12]. This will minimize the error estimate of the late reflection power and further improve the Wiener gain for effective wavelet filtering.

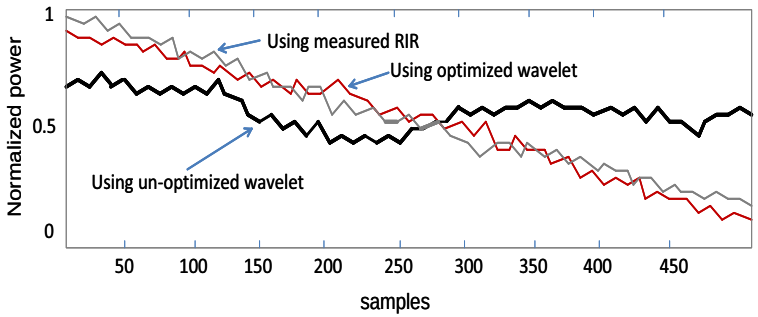

図 4 Normalized late reflection power plot.

4. Experimental Evaluation

The training database is from the Japanese Newspaper Article Sentence (JNAS) corpus. The open test set is composed of 200 utterances. ASR experiments are carried out on the Japanese dictation task with 20K vocabulary. The language model is a standard word trigram model. The acoustic model is a phonetically tied mixture (PTM) HMMs with 8256 Gaussians in total. We experimented in the condition of reverberation time: T_{60}=200 msec and 600 msec. Reverberant training data are synthetically produced with the automatically generated RIR discussed in Section 3.1. The test data were recorded in a room with known reverberation time: T_{60}=200 msec and 600 msec. Thus, we used actual reverberant data for evaluation. In the experiments we used a total number of bands $M = 5$ which was found to be effective[1)2)]. The wavelet used here is the Daubechies wavelet which was also used in[12)].

In Table 1, we show the ASR performance in word accuracy for different methods. (A) and (B) are the results when the reverberant data are not processed and matched against clean and reverberant acoustic models, respectively. We show the result of an approach based on improved wavelet thresholding[10)] in (C). This method is an improvement of the simple thresholding in[9)]. By incorporating additional information such as VAD and statistical profile of the contaminant data (i.e. reverberation), an improved thresholding is achieved. In (D), we show an improvement of the performance from (C) when the wavelet parameters are optimized as proposed in Section 3.1. Another wavelet-based dereverberation method based on extrema clustering[11)] is shown in (E). This method adopts the speech production model to detect the reverberant coefficients. It applies wavelet extrema clustering to the linear prediction coefficients to separate the clean and reverberant components. When the wavelet parameters are optimized, the recognition performance is further improved in (F). The result of our previous dereverberation approach[12)] is shown in (G) and the result of incorporating wavelet optimization is given in (H). The results in Table 1 show the effect of optimizing the wavelet parameters in the recognition performance. The consistent improvement is observed across the different wavelet-based methods.

In Fig. 4, we show the power plot of the late reflection, estimated for both optimized and un-optimized wavelet parameters. We also show the exact power by reproducing the exact late reflection using the measured RIR. In this plot, the power envelope when using the optimized wavelet parameters closely resembles that of the exact late reflection power (using measured RIR). This suggests, that the optimized wavelet is able to track the existence of late reflection power in a reverberant signal. We note that the reverberant signal contains speech energy as well. The estimation when using un-optimized wavelet is not good as it cannot discriminate properly between the clean speech and the late reflection in the reverberant signal.

5. Conclusion

We have proposed an improved dereverberation approach based on wavelet filtering. By optimizing the wavelet parameters, the system can effectively estimate the power of the clean speech and the late reflection in a reverberant signal. This results to an effective Wiener gain for dereverberation. Most of the processes in the dereverberation scheme are closely linked to the acoustic model likelihood. Thus, the proposed dereverberation method is effective in achieving robustness in the ASR application.

参 考 文 献

1) R. Gomez et.al. , “Distant-talking Robust Speech Recognition Using Late Reflection Components of Room Impulse Response” *ICASSP*, 2008

2) R. Gomez, T. Kawahara, “Optimization of Dereverberation Parameters based on Likelihood of Speech Recognizer” *Interspeech*, 2009.
3) K. Kinoshita , T. Nakatani and M. Miyoshi, “Spectral Subtraction Steered By Multi-step Forward Linear Prediction For Single Channel Speech Dereverberation” *ICASSP*, 2006
4) R. Gomez, J. Even, H. Saruwatari and K. Shikano, “Fast Dereverberation for Hands-Free Speech Recognition” *IEEE Workshop HSCMA*, 2008
5) H.-G. Hirsch and H. Finster, “A new approach for the adaptation of HMMs to reverberation and background noise” *Speech Communication*, pp 244-263, 2008.
6) S. Ayat et,al., “An Improved Wavelet-based Speech enhancement by Using Speech Signal Features” *Computers and Electrical Engineering* 2006.
7) E. Ambikairajah et. al., “Wavelet Transform-based Speech Enhancement” *ICSLP*, 1998
8) Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, ”An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses”
9) D.L. Donoho, “Denoising by soft thresholding”, *IEEE Trans. Info. Theory* 1995.
10) H. Sheikhzadeh and Hamid. Abutalebi, “An Improved Wavelet-based Speech Enhancement System” *Eurospeech*, 2001.
11) S. Griebel and M. Brandstein, “Wavelet Transform Etxrema Clustering for Multi-channel Speech Dereverberation”
12) R. Gomez and T. Kawahara , “Optimizing Spectral Subtraction and Wiener Filtering for Robust Speech Recognition in Reverberant and Noisy Conditions” *ICASSP*, 2010