Robust Front End Processing for Speech Recognition in Reverberant Environments: Utilization of Speech Characteristics

Rico Petrick¹, Xugang Lu², Masashi Unoki², Masato Akagi², Ruediger Hoffmann¹

¹Laboratory of Acoustics and Speech Communication, Dresden University of Technology, Germany
²School of Information Science, Japan Advanced Institute of Science and Technology, Japan

[© the speaker-microphone-distance (SMD) microphone. The direct sound energy (field) feature analysis, temporal power envelope feature analysis is tested, which is another new contributions in this paper. Their performance considering the speaker-to-microphone distance is tested, which is another new contributions in this paper. Their performance in dependency on the reverberation time, SMDs in a SMART room [2] at UPC in Barcelona.]

Abstract

This paper proposes two methods for robust automatic speech recognition (ASR) in reverberant environments. Unlike other methods which mostly apply inverse filtering by blindly estimated room impulse responses to achieve dereverberation, the proposed methods are based on the utilization of the characteristics of speech. The first method - Harmonicity based Feature Analysis – takes advantage of the harmonic components of speech, which are assumed to be undistorted. The second method - Temporal Power Envelope Feature Analysis – utilizes the temporal modulation structure of speech, representing the phoneme level temporal events which contain most intelligibility information. Both methods increase the recognition performance remarkably in a different way. Combining both of them connects their individual advantages. In order to examine the performance of utilizing harmonicity and modulation temporal structure for reverberant ASR, the methods are tested in clean and reverberant training. As results show, even in strong reverberant conditions both methods obtain practical applicable performance for reverberant training. In addition, besides testing their performance in dependency on the reverberation time, their performance considering the speaker-to-microphone distance is tested, which is another new contributions in this paper. Index Terms: reverberation, robust ASR, harmonicity based feature analysis, temporal power envelope feature analysis

1. Introduction

Reverberation is one of the major and still unsolved problems of current research on automatic speech recognition (ASR). It has a strong degrading effect on the recognition rate (RR) [1]. Methods in the more traditional field of noise robustness are not applicable since reverberation and noise have different effects. Effects of room acoustics: Inside of rooms reverberation smears the spectro-temporal structure of speech. The reverberant signal \( x(t) \) consists of direct (clean) and reverberant (disturbance) sound components \( (x_D(t) \text{ and } x_R(t)) \) which add at the microphone. The direct sound energy \( (E_D) \) degrades with the speaker-microphone-distance (SMD) \( d \) following \( E_D(d) \approx 1/d^2 \). The reverberation sound energy (field) \( (E_R) \) is position independent \( (E_R(d) \approx \text{const.}; \text{ideal assumption}) \). Both result in a position dependent signal to reverberation ratio \( (\text{SRR}(d)) \), decreasing with the SMD. The SMD, where \( w_D = w_R \) is known as the reverberation distance \( r \). Far (approx.: \( \text{SMD} > r \)); \( w_D < w_R \) and near (approx.: \( \text{SMD} < r \)); \( w_D > w_R \) sound field behavior can be distinguished. The system between the speaker and microphone can be described by the room impulse response (RIR) \( h(t) \). This also contains direct \( (h_D(t), \text{impulse}) \) and reverberation \( (h_R(t), \text{tail}) \) components. The energy of \( h_R(t) \) decays exponentially (ideal assumption), which leads to a linear function in dB, decaying with the velocity \(-60 \text{dB}/\text{T}_{60} \). The reverberation time \( T_{60} \) is the most commonly used property to describe a specific room. Hence, many ASR and dereverberation researchers have utilized \( T_{60} \) as the only parameter to evaluate the quality of their systems in reverberant environments. Figure 1 shows a three-dimensional graphic of the SCHROEDER-integral of RIRs in decibels \( (L_{RIR}(t, r)) \) measured at varying SMDs. It gives a graphical explanation why the only use of \( T_{60} \) is insufficient by far to describe dependencies of systems on reverberation; it is only suitable for the far sound field. The investigation of the dependency on the SMD is a new contribution of the paper. Inverse-filtering-based dereverberation: Traditional and new approaches are proposed to solve the dereverberation task (e.g., [3, 4, 5]). Most dereverberation algorithms are based on blind estimation of the RIR and subsequent filtering of the input signal with the inverse of \( h(t) \). However, certain blind estimation of RIRs is a tough task, which is even more problematic while tracking of changing RIRs (varying speakers location, moving objects/persons inside the room). Long reverberation times lead to more unstable systems. Adaptation times are mostly far too long to enable practical applications for command-word ASR.

This work was supported by a Grant-in-Aid for Scientific Research (No. 18680017) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan, and by the Foundation of German Business (Stiftung der Deutschen Wirtschaft (sdw)). It was also partially supported by the SCOPE (071705001) of MIC, Japan. Measurements of the Room Impulse Responses in the SMART room where accomplished during a midterm visiting research period at the TALP Research Center in the UPC Barcelona in collaboration with Carlos Segura and under the coordination of Prof. Climent Nadeu.

---

Figure 1: SCHROEDER-integral of RIRs, measured at different SMDs in a SMART room [2] at UPC in Barcelona.
Dereverberation utilizing speech characteristics: Approaches that do not need to estimate the RIR are preferred for real ASR systems. Unlike signal enhancement systems which demands sophisticated speech quality, ASR has the advantage, that only features need to be restored. One possible way of enhancing features is the utilization of speech properties, which are not affected by reverberation. The authors propose two techniques, each utilizing a different speech property: (I) Harmonicity-based feature analysis (HFA) [6] utilizes two assumptions of reverberant speech: (a) harmonic components are assumed to be undisturbed and (b) the low frequency energy of unvoiced sections is most probably reverberation and can be deleted. (II) Temporal power envelope feature analysis (TPEFA) as similar in [7, 8] utilizes the temporal modulation structure that remains under reverberant conditions.

Requirements for applicable command word ASR: Applicable ASR in reverberant rooms requires: (a) a robust acceptable recognition rate (RR) ($\approx 90\%$ [9]) under typical varying indoor conditions (living/home/office environments: $0.3 \, \text{s} < T_{90} < 1.0 \, \text{s}; 0.5 \, \text{m} < \text{SMD} < 4 \, \text{m}$), (b) no or real time adaptation ($< 2.0 \, \text{s}$), (c) robustness against changes in the RIR (movements of speakers/object) and (d) feasible numerical complexity. Practical applicable dereverberation methods have to meet these demanding requirements.

2. Utilizing Speech Properties for Rev. ASR

In this paper only a brief overview of implemented ideas of the applied front-end methods is given. The exact implementations are described in associated references: (a) CFA (Conventional Feature Analysis) as used in [10], (b) HFA proposed in detail in [6], (c) TPEFA as used in [7, 8] and (d) HFA+TPEFA combined proposed in this article. Before any of these front-end methods are evaluated, the model has to be trained applying the appropriate front-end on the training data.

2.1. Conventional Feature Analysis

For comparison subsequent methods have to adapt to the same conditions as the CFA used in [1, 10]. These involve classical short time Fourier analysis (STFA) followed by a mel filterbank (MFB). STFA includes framing of the input signal $x(k)$ ($f_s = 16 \, \text{kHz}$) into $x(a, k)$ (frame index $a$), windowing and FFT ($N = 512$). MFB includes a logarithm and cepstral smoothing on magnitude spectrum $X(a, n)$, energy derivation and normalization of 30 mel-scaled filter channels composing a feature vector, $\vec{f}(a)$. The absolute frame energy is added as the 31st component.

2.2. Harmonicity-based Feature Analysis

HFA implements three ideas:

(i) Harmonic components are assumed to be clean: This principle is already used in [5]. HFA synthesizes voiced spectra $X_{v,v}(n)$ based on the measured harmonic components at harmonic indices $n_h$ (multiples of $F_0$). Waveform interpolation is carried out between two $n_h$‘s taking into consideration the typical structure of logarithmized voiced spectra.

(ii) Unvoiced speech is highly reverberated at low frequencies: Unvoiced speech sections, e.g., fricatives, have their main features in the higher frequency regions. Their lower frequency regions are highly distorted by reverberation coming from previous voiced sections as shown in Fig. 2. These low frequency reverberation have high energy compared to the unvoiced features due to the more energetic production process of voiced speech. Therefore HFA barely suppresses low frequency components, enhancing the structure of the feature vector into a more unvoiced shape. Some information is lost for unvoiced wideband signals, but this also applies to the training data. This processing also recovers the low frequency temporal structure of speech, which is actually the key issue in TPEFA.

(iii) High frequency reverberation is harmless: According to [6] reverberation above 2500 Hz is almost harmless for ASR. Therefore, HFA involves the previous ideas of (i) voiced and (ii) unvoiced speech only at low frequencies. High frequency components remain unchanged. The fading interaction between the original high frequency components and the two types of low frequency processes is smoothly accomplished by a spectral overlap-and-add. A number of experiments achieved optimal fading parameters [6].

The implementation follows Fig. 3(a). As previously pointed out, this behavior results in two different types of analysis for voiced and unvoiced frames, generating the synthetic spectra, $X_{v,v}(n)$ and $X_{s,u}(n)$, which are passed to the MFB.

$F_0$ estimation: Initially the autocorrelation function (ACF) method is used. Under reverberation this simple approach performs similar compared to advanced methods [11].

VUD: Voiced unvoiced decision (VUD) also uses a simple approach where the mean energy of the harmonic components of a frame is compared to a dynamically derived threshold [6].

Considering $F_0$ estimation and VUD errors: Using other more sophisticated approaches for $F_0$ estimation, $F_0$ post-processing or VUD (e.g., [12]) did not lead to better results of RR, concluding that these easy approaches appeared as sufficient. Their errors are handled by the model, since errors also occur while analyzing the training data. Error modeling performs even better for reverberant training (compare the results in sections 3.1 and 3.2). This is especially the case for VUD errors, which therefore demand at least a two-Gaussian Mixture Model (two-GMM) Hidden Markov Model (HMM). Due to occurring VUD errors, two different methods of analysis generating $X_{v,v}(n)$ and $X_{s,u}(n)$ can be undertaken for the same phoneme (one for the correct and the second for the incorrect VUD), forming two distant clusters in the feature space for the same phoneme. One could argue that analyzing voiced frames in an unvoiced manner would delete too much information for discrimination. Only low frequency components are suppressed but second and third formant information still remains resulting in slightly reduced discrimination. Despite these errors, the overall processing of HFA increases the performance of the ASR for reverberant training even more.

2.3. Temporal Power Envelope Feature Analysis

Recent research shows that most speech intelligibility information is encoded in the temporal modulation envelopes (TMEs) of frequency subbands [13]. Furthermore, these TMEs are robust against noise distortion in speech-enhancement systems, they have to be restored. Following the same idea to enhance ASR features, techniques such as Relative Spectral Filtering (RASTA) for spectral or cepstral trajectories [14] are proposed.
For the reverberant conditions the authors investigate and discuss, that the fine temporal structure of speech is smeared but the large-scale TME structures corresponding to linguistic events (phonemes) are still retained. However, they are flattened due to reverberation. As some researchers (e.g., [15]) show, most speech intelligibility information is distributed in the TME structures between 2 Hz and 20 Hz. Concluding, occurring higher modulation frequencies can be seen as induced by distortions as reverberation. According to these assumptions, the implemented TPEFA front-end (Fig. 3(b)) aims to restore the temporal modulated power envelope (TPE) for frequencies below 20 Hz. Considering the temporal co-modulation property of speech [7], \( x(k) \) is decomposed into \( C = 64 \) evenly distributed frequency bands (channel index \( c \)), time domain constant band filterbank (CBFB), and a bandwidth of 100 Hz according to previous research [7, 8] by the authors. Each subband signal \( x_c(k) \) can be regarded as a temporal (amplitude) modulated signal:

\[
x_c(n) = \tilde{\alpha}_c(n) \cos \left( \omega_c \frac{k}{T_a} + \tilde{\varphi}_c \right),
\]

where \( \tilde{\alpha}_c, \tilde{\varphi}_c \) are the TME, \( \omega_c \) and phase of the \( c \)-th subband. To extract TPE \( \tilde{\alpha}_c(k) \), the squared magnitude of the complex analytical signal \( z_c(k) \) is derived. \( z_c(k) \) is composed of \( x_c(k) \) as the real part and the Hilbert transform \( \text{Hilbert}(\cdot) \) of \( x_c(k) \) as the imaginary part. Subsequently, \( \tilde{\alpha}_c(k) \) is low-pass filtered (LPF with a cutoff frequency of 20 Hz according to [15], ref. above):

\[
\tilde{\alpha}_c(k) = \text{LPF} \left[ |x_c(k) + j\text{Hilbert}[x_c(k)]|^2 \right].
\]

\( \tilde{\alpha}_c(k) \) is still a time signal. The index \( c \) can be seen as a frequency index of a \( C \)-channel TPE spectrum for each time index \( k \). To use this preprocessor for ASR, framing is applied by down sampling with \( M_k = L_k \) (frame interval \( T_a = 160 \), the same as for CFA) resulting in \( \tilde{\alpha}_c(a) \). No anti-aliasing filter is needed, because of the previous LPF (20 < \( f_s/(2L_k) \) Hz). To compare the performance of TPEFA with the other methods, \( \tilde{\alpha}_c(a) \) is interfaced with MFB, which requires \( N/2 \) frequency bins as input. Therefore, \( \tilde{\alpha}_c(a) \) is up-sampled in frequency by \( L_e = 4 \) (simple trapezoid interpolation).

### 2.4. Combination of HFA and TPEFA

HFA+TPEFA uses the synthesized spectra \( X_c(a, n) \) generated within HFA for resynthesis into short time signals \( x_c(a, k) \) applying the Fourier series, which incorporates the original phase \( \varphi(X_c(a, n)) \). An overlap-and-add algorithm assembles the HFA-processed time signal as input for the TPEFA. An inverse processing order, i.e., TPEFA before HFA is not possible since TPEs cannot be resynthesized or retain harmonicity information. For the same reason HFA also HFA+TPEFA requires at least a 2-GMM model.
particular training condition, but drops the RR for other conditions. HFA increases the RR of CFA and keeps it stable under various rev. conditions when training becomes more reverberant. This can especially be observed in Fig. 5(d1) (HFA compared to TPEFA and CFA). TPEFA performs significantly better than CFA but also better than HFA (in most cases). However, it tends to support the actual training conditions and decreases under different test conditions. HFA+TPEFA again combines the advantages of HFA and TPEFA (stable vs. high improvement). $T_{60,\text{Train}} = 1$ s achieves the best overall results (both top and bottom figures should always be considered for rating).

4. Conclusions

Comprehensive recognition experiments show that both applied methods, HFA and TPEFA, can improve recognition. Although the RR is drastically increased (e.g., from 35% up to 70%) under clean training conditions, the performance is still insufficient for practical considerations (< 90%). Additional reverberant training achieves practical application requirements; also for varying reverberation conditions. The gain at HFA is caused by harmonic components, which can be considered as clean and by the deletion of low frequency reverberation in unvoiced speech, which is highly disturbing. HFA suffers at clean conditions since some information is deleted but stably improves in reverberant conditions. TPEFA gains the ASR performance by information about the temporal envelope modulation, which is a robust information carried by speech also in noisy and reverberant environments. However TPEFA tends to support the applied training condition. The combination HFA+TPEFA takes advantages of both methods (stable improvement and high improvement) and compensates their weak points. The enhancement of both methods is achieved by emphasizing feature information by characteristic preferences of speech, resulting in high practical applicable RRs even in adverse environments. No adaptation as in current dereverberation approaches is needed leading to real time processing ability, required in command word recognition applications. A disadvantage of the TPEFA is a high processing load due to the time domain filter bank, which cannot be handled by current embedded devices, but will be in future systems.

5. References


Figure 5: Dependency of the RR on $T_{60,\text{Test}}$ in near field (SMD$_{\text{Test}} = 1$ m, top diagrams) and in far field (SMD$_{\text{Test}} = 3$ m, bottom diagrams). Applied training conditions: $T_{60,\text{Train}} = (a)$ 0 s (clean), (b) 0.4 s, (c) 1.0 s and (d) 2.0 s at SMD$_{\text{Train}} = 1$ m.