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

Rico Petrick[1], *Xugang Lu*[2], *Masashi Unoki*[2], *Masato Akagi*[2], *Ruediger Hoffmann*[1]

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

[Rico.Petrick,Ruediger.Hoffmann]@ias.et.tu-dresden.de [xugang,unoki,akagi]@jaist.ac.jp

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 ($x_D(t)$ and $x_R(t)$) which add at the microphone. The direct sound energy (field) w_D degrades with the speaker-microphone-distance (SMD) r following $w_D(r) \approx 1/r^2$. The reverberant sound energy (field) is position independent ($w_R(r) \approx$ const.; ideal assumption). Both result in a position dependent signal to reverberation ratio ($\mathrm{SRR}(r)$, decreasing with the SMD. The SMD, where $w_D = w_R$ is known as the reverberation distance r_R. Far (approx.: SMD $> r_R$; $w_D < w_R$) and near (approx.: SMD $< r_R$; $w_D > w_R$) 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 ($h_D(t)$, impulse) and reverberation ($h_R(t)$, tail) components. The energy of $h_R(t)$ decays exponentially (ideal assumption), which leads to a linear function in dB, decaying with the velocity $-60\,\mathrm{dB}/T_{60}$. 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 ($L_{\mathrm{Schr}}(t, r)$) 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.

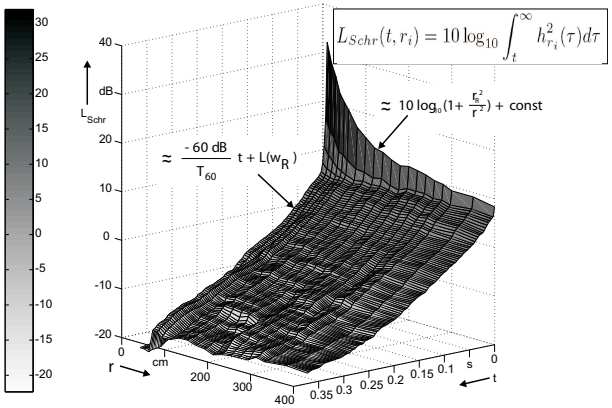

Figure 1: SCHROEDER-*integral of RIRs, measured at different SMDs in a SMART room [2] at UPC in Barcelona.*

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.

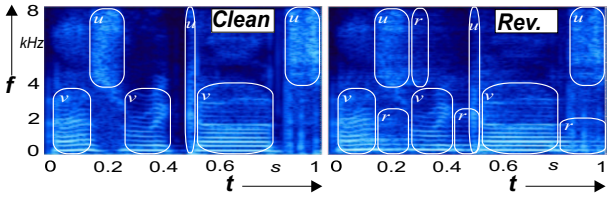

Figure 2: *Schematic illustration of disturbances caused by reverberation (r) for voiced (v) and unvoiced (u) speech.*

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 s $< T_{60} <$ 1.0 s; 0.5 m $<$ SMD $<$ 4 m), (b) no or real time adaptation ($<$ 2.0 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 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, $\vec{x}(a)$. 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 $X_{s,v}(n)$ 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, $X_{s,v}(n)$ 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 postprocessing 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 $X_{s,v}(n)$ 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 researches show that most speech intelligibility information is encoded in the temporal modulation envelopes (TMEs) of frequency subbands [13]. Furthermore, as 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.

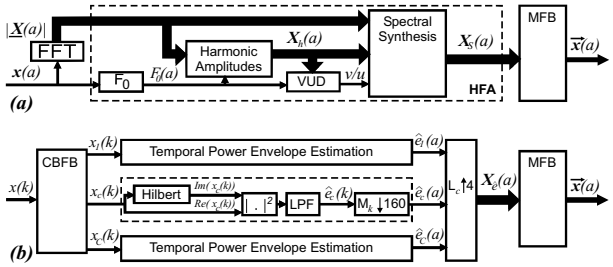

Figure 3: *Block diagrams of (a) HFA and (b) TPEFA.*

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, occuring 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:

$$x_c(n) = \hat{\alpha}_c(n) \cos\left(\omega_c \frac{k}{f_s} + \varphi_c\right), \qquad (1)$$

where $\hat{\alpha}_c$ is the TME, ω_c and φ_c are the associated carrier frequency and phase of the c-th subband. To extract TPE $\hat{e}_c(k)$, the squared magnitude of the complex analytical signal $\underline{x}_c(k)$ is derived. $\underline{x}_c(k)$ is composed of $x_c(k)$ as the real part and the Hilbert transform (**Hilbert** $[\cdot]$) of $x_c(k)$ as the imaginary part. Subsequently, $\hat{e}_c(k)$ is low-pass filtered (**LPF** with a cutoff frequency of 20 Hz according to [15], ref. above):

$$\hat{e}_c(k) = \mathbf{LPF}\left[|x_c(k) + j\mathbf{Hilbert}\left[x_c(k)\right]|^2\right]. \qquad (2)$$

$\hat{e}_c(k)$ 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 $M_k = I_a$ (frame interval $I_a = 160$, the same as for CFA) resulting in $\hat{e}_c(a)$. No anti-aliasing filter is needed, because of the previous **LPF** ($20 < f_s/(2I_a)$ Hz). To compare the performance of TPEFA with the other methods, $\hat{e}_c(a)$ is interfaced with MFB, which requires $N/2$ frequency bins as input. Therefore, $\hat{e}_c(a)$ is up-sampled in frequency by $L_c = 4$ (simple trapezoid interpolation).

2.4. Combination of HFA and TPEFA

HFA+TPEFA uses the synthesized spectra $X_s(a, n)$ generated within HFA for resynthesis into short time signals $x_s(a, k)$ applying the Fourier series, which incorporates the original phase $\varphi(\underline{X}(a, n))$. 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 as HFA also HFA+TPEFA requires at least a 2-GMM model.

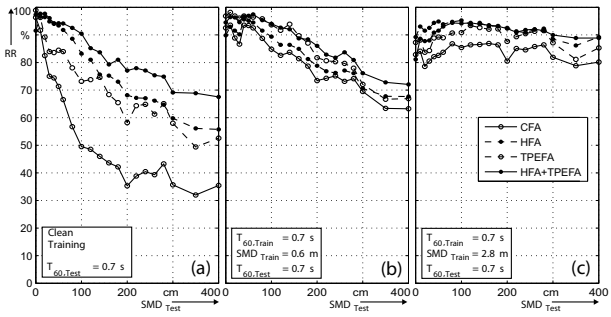

Figure 4: *RR dependent on SMD_{Test}. Training conditions: SMD_{Train} = (a) 0.0 m (clean), (b) 0.6 m and (c) 2.8 m.*

3. Evaluation

This work uses the same evaluation system as described in [1, 6] (UASR recognizer, subset of APOLLO corpus [9] consisting of 1020 command phrases (each ≈2 s speech) of 17 classes. Two sets of dependency experiments are accomplished:
•**RR(SMD)** in the SMART room environment (Fig. 1)
•**RR(T_{60})** in rooms for near and far field (SMD = 1 m/3 m).

3.1. Evaluation Results for Clean Training

•**RR(SMD)** (Fig. 4 *(a)*): A strong degradation using CFA can be observed even after a few cm of SMD. HFA and TPEFA increase the performance over the whole range, closely won by HFA. HFA+TPEFA performs best and takes advantages of both methods. Interesting point: The results adumbrate the reverberation distance of this room ($r_{\mathrm{R}} \approx$1.5,...,2.5 m).
•**RR(T_{60})** (Fig. 5 *(a1)*, *(a2)* (near, far field)): CFA again performs poorly, decreasing with increasing $T_{60,\mathrm{Test}}$. HFA gradually improves for clean training, due to some loss of information in the undistorted data (reverberant training compensates for this effect, as described below). For increasing $T_{60,\mathrm{Test}}$, the degradation in RR is less than for CFA; → HFA increases the RR for reverberant conditions. TPEFA enhances the general information properties of speech, increasing the RR already for the clean case but also over the whole reverberant test range. HFA+TPEFA leads to a slight drop at clean conditions compared to TPEFA, due to the loss caused by HFA. But for the more reverberant conditions, the advances of HFA and TPEFA add again.

3.2. Evaluation Results for Reverberant Training

In difference to speech enhancement systems, ASR has the advantage to train models at the disturbing conditions. However, a dedicated reverberant model usually tends to support the training condition, but drops other conditions (refer CFA in all diagrams). Good behavior is achieved by a method when a training condition can be used for general test conditions.
•**RR(SMD)** (Fig. 4 *(b)*, *(c)*): The model is trained under reverberant conditions of the SMART room at several SMDs. CFA performs better for the far field, but loses RR in the near field. HFA training at $\mathrm{SMD} = 280$ cm achieves the best performance, although there is a slight decrease for clean case. TPEFA performs significantly better than CFA, but the results tend to support the training condition resulting in a loss for clean test data. HFA+TPEFA is marginally outperformed by HFA, due to the RR drop of TPEFA at short SMDs.
•**RR(T_{60})**(Fig. 5 *(b) – (d)*): Reverberant training conditions at several T_{60}'s (at $\mathrm{SMD} = 1$ m; far field training (3 m) did not lead to good results) are applied. CFA performs better for the

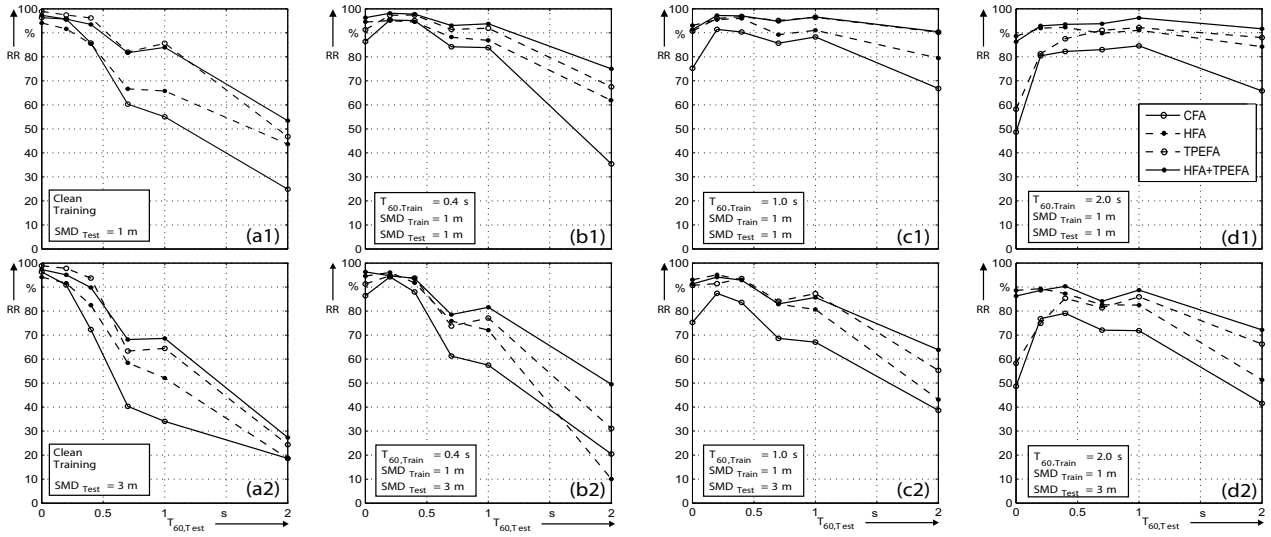

Figure 5: *Dependency of the RR on $T_{60,\mathrm{Test}}$ in near field (SMD_{Test} =1 m, top diagrams) and in far field (SMD_{Test} =3 m, bottom diagrams). Applied training conditions: $T_{60,\mathrm{Train}}$ = (a) 0.0 s (clean), (b) 0.4 s, (c) 1.0 s and (d) 2.0 s at SMD_{Train} =1 m.*

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). $T_{60,\mathrm{Train}} = 1$ 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 condition. 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

[1] Petrick, R., Lohde, K., Wolff, M. and Hoffmann, R., "The harming part of room acoustics for automatic speech recognition," *Proc. INTERSPEECH2007*, pp. 1094–1097, Antwerp, 2007.

[2] Neumann, J., Gasas, J. R., Macho, D., Hidalgo, J. R., "Integration of audio-visual sensors and technologies in a smart room," *Personal and Ubiquitous Computing*, Springer London, ISSN: 1617-4909 (print), 1617-4917 (online), April 2007.

[3] Miyoshi, M. and Kaneda, Y., "Inverse filtering of room acoustics," *IEEE Trans. ASSP*(36), pp. 145–152, 1988.

[4] Gillespie, B. W., Malvar, H. S., and Florencio, D. A., "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Proc. ICASSP2001*, pp. 3701–3704, Salt Lake City, 2001.

[5] Nakatani, T., Kinoshita, K., and Mihyoshi, M., "Harmonicity based blind dereverberation for single-channel speech signals," *IEEE Trans. ASLP* 15(1), pp. 80–95, 2007.

[6] Petrick, R., Lohde, K., Lorenz, M., and Hoffmann, R., "A new feature analysis method for robust ASR in reverberant environments based on the harmonic structure of speech," *Proc. EUSIPCO2008*, Lausanne, 2008 (accepted).

[7] Unoki, M., Sakata, K., Furukawa, M., and Akagi, M., "A speech dereverberation method based on the MTF concept in power envelope restoration," *Acoust. Sci. & Tech.*, pp. 243–254, 25(4), 2004.

[8] Lu, X., Unoki, M., and Akagi, M., "Comparative evaluation of modulation-transfer-function-based blind restoration of sub-band power envelopes of speech as a front-end processor for automatic speech recognition systems," *Acoust. Sci. & Tech.*, 2008 (in press).

[9] Maase, J., Hirschfeld, D., Koloska, U., Westfeld, T., and Helbig, J., "Towards an evaluation standard for speech control concepts in real-world scenarios," *Proc. EUROSPEECH2003*, pp. 1553–1556, Geneva, 2003.

[10] Hoffmann, R., Eichner, M. and Wolff, M. "Analysis of verbal and nonverbal acoustic signals with the Dresden UASR system," In Esposito, A., et al. (eds.): *Verbal and Nonverbal Communication Behaviours*, pp. 200–218, Berlin etc.: Springer, LNAI 4775, 2007.

[11] Petrick, R., Unoki, M., Mittal, A., Segura, C., and Hoffmann, R., "A comprehensive study on the effects of room reverberation on fundamental frequency estimation," *Proc. INTERSPEECH2008*, Brisbane, 2008 (accepted).

[12] Luengo, I., Saratxaga, I., Navas, E., Hermaez, I., Sanchez, and Sainz, J., "Evaluation of pitch detection algorithms under real condition," *Proc. ICASSP2007*, pp. 1057–1060, Honolulu, 2007.

[13] Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," *Science*, 270, pp. 303–304, 1995.

[14] Hermansky, H., Morgan, N., and Hirsch, H. G., "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *ICASSP'93*, pp. 83–86, 1993.

[15] Dau, T., "Modeling auditory processing of amplitude modulation," Ph.D. Thesis, ISBN 3-8142-0570-7, 1996.