16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

A NEW FEATURE ANALYSIS METHOD FOR ROBUST ASR IN REVERBERANT ENVIRONMENTS BASED ON THE HARMONIC STRUCTURE OF SPEECH

Rico Petrick[1] , *Kevin Lohde*[2] , *Mike Lorenz*[1] *and Ruediger Hoffmann*[1]

[1]Dresden University of Technology, Laboratory of Acoustics and Speech Communication
[2]Dresden University of Applied Sciences, Department of Electrical Engineering
[Rico.Petrick,Ruediger.Hoffmann]@ias.et.tu-dresden.de

ABSTRACT

This article proposes a new signal analysis method for automatic speech recognition designed to aim high robustness against distortions caused by room reverberation. The method is initially named Harmonicity based Feature Analysis (HFA) and implements the following three ideas: (i) reconstruction of a spectrum from the harmonic components (assumed to be undistorted) of a voiced speech spectrum. (ii) suppression of disturbing reverberation in unvoiced spectra coming from previous voiced sections. (iii) high frequency regions are not affected by HFA since they have negligible effect on the recognition rate. HFA works on the basis of fundamental frequency estimation and voiced/unvoiced decision. Evaluation results show significant improvement of the recognition performance over a wide range of reverberant conditions while using HFA in connection with reverberant training. Apart from good performance, advantages of HFA compared to state of the art dereverberation approaches are real time processing (no adaptation time) and robustness against changes of the room impulse response.

1. INTRODUCTION

The performance of automatic speech recognition (ASR) systems in real environments suffers from the mismatch between the acoustic training and test conditions. Disturbances can be divided into additive (noise) and convolutive distortions (reverberation). In the last three decades a number of methods have been developed, to deal with disturbances caused by additive noise. By contrast research for methods against reverberation is a rather new and still a challenging task . Methods dealing with noises cannot be adapted for reverberation, because of the very different behavior of the two distortion types. Conventional feature analysis (CFA) methods, e.g. Mel-Filter-Bank (MFB) or Mel-Frequency-Cepstral-Coefficients (MFCC) fail in the ambiance of reverberation (refer "baseline" in Figure 8). Former investigations e.g. [1] have observed the strong degrading effect on the recognition rate (RR) of ASR systems due to reverberation disturbances. The training of the ASR system with reverberant training data (multi condition training (MCT)) may be a solution for a dedicated room environment. However, the results in section 3 show that CFA with MCT is not sufficient for varying reverberation conditions. Different methods (including very promising approaches as [2, 3]) have been introduced to handle the problem in the signal domain with blind dereverberation algorithms. Some of them need high adaptation times (e.g. [4, 5]), others can suffer from changes of the room impulse response (RIR) due to moving speakers. Also feature and model based approaches have been introduced (e.g. [6, 7]). This article proposes the method Harmonicity based Feature Analysis which works in the signal (spectral) domain. HFA is designed to meet the following requirements, which are demanded in a real world ASR system (application) at typical indoor conditions:

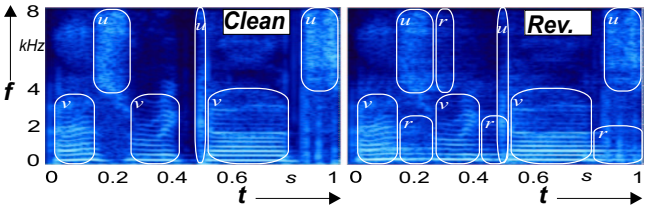

Figure 1: *Schematical illustration of disturbances caused by reverberation (r) for voiced (v) and unvoiced (u) speech. Left: clean signal. Right: reverberant signal ($T_{60} = 0.4$ s ; SMD = 100 cm). Remark: The rather low reverberant condition in this figure is chosen for illustration reasons of the principle. Stronger reverberation (e.g. $T_{60} = 0.7$ ... 2 s) have a much worse effect*

- robust high RR under varying reverberation conditions (varying reverberation time T_{60} and speaker microphone distances *SMD*)
- No adaptation or adaptation in real time
- Coverage of typical indoor conditions (living/office environments: $T_{60} = 300 \ldots 1000$ ms; $SMD = 0.5 \ldots 4$ m)
- Robustness to changes of the speaker position.
- Feasible numerical complexity.[1]

Current proposed methods mostly fail in at least one of these requirements.

2. PROPOSED METHOD: HFA - HARMONICITY BASED FEATURE ANALYSIS

2.1 Motivation

This section gives a rough overview about the ideas used in HFA. The clear mathematical description is given in the sections below. HFA is motivated by three ideas:

(i) Harmonic components are assumed as clean: The basic idea of HFA was inspired by the work of Nakatani et al. (first published in [5]; extensive publication [8]). Nakatani et al.'s method HERB - Harmonicity-based-dEReverBeration - approaches the blind dereverberation problem by system inversion. The blindly estimated inverse system is a long time average of divisions of the spectra of a synthesized signal $\hat{x}(t)$ and the reverberant speech

[1]Applies for state of the art computers. But even for increasing computer abilities this may be an issue for embedded solutions in the future

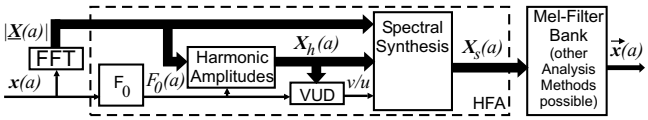

Figure 2: *Block diagram of the proposed method HFA*

signal $x'(t)$. $\hat{x}(t)$ is derived from the harmonic components of $x'(t)$. The idea is that the harmonic spectral components are far above of the noise and reverberation level, so they can be assumed as direct (clean) components of $x'(t)$. The synthesis is achieved by a frame based Fourier series of the harmonic spectral components followed by an overlap-and-add method. The main disadvantage of HERB is the long adaptation time due to long time averaging, which is far away from real time processing. The idea of HFA is the use of the "nonreverberant" signal $\hat{x}(t)$ for the recognition of speech. In difference to Nakatani et al.'s approach HFA does not apply the real synthesis of the signal. Instead HFA generates a sequence of "nonreverberant" spectra, which are reconstructed (synthesized) only from the harmonic components of the reverberant signal $x'(t)$. This processing is due to voiced frames only.

(ii) Unvoiced speech is highly reverberated due to previous voiced speech: Clean unvoiced speech sections, e.g. fricatives, have their main features in the higher frequency regions. As shown in Figure 1, in ambiance of reverberation the lower frequency regions of unvoiced sections are highly distorted by reverberation coming from previous voiced sections. Voiced speech sections have more energy than unvoiced sections due to their different production processes. That means, that also these distortions have much more energy than the actual fricative, which leads to a complete mismatch for a clean trained model of an unvoiced phoneme. In fact such a frame would be rather recognized as one of the voiced phonemes. To reconstruct some of the structure components of an unvoiced spectrum, the frequencies belonging to reverberation from previous voiced sections are suppressed in unvoiced frames. If the unvoiced component is a wide band signal (like a plosive), its low frequency features will be eliminated. But if this also applies to the training data, equal training and test conditions are achieved, although some information is lost.

(iii) High frequency reverberation is harmless: The investigations in [1] have clearly shown, that high frequency reverberation above about 2500 Hz has almost no degrading effect on the RR. Contrarily a rather enhancing behavior on the RR for high frequency reverberation was found in [1] . This effect is used in HFA. The two previous described procedures for voiced and unvoiced frames **(i)** and **(ii)** are only applied at low frequency components (approx. $<$ 2500 Hz). High frequency components are left unchanged.

Before applying the HFA front end for recognition, the recognizer has to be trained with HFA processed training data.

2.2 Method Overview

HFA is embedded as signal processing module in the feature analysis of an ASR system (Ref. Figure 2). The sampling frequency amounts $f_s = 16$ kHz. $x(a,k)$ is introduced as a sequence of the framed input signal in the time domain with the frame index a ($0 \leq a < A-1$) and the frame interval $I_a = 160$ ($\hat{=}$ 10 ms). k refers to the frame internal sample index ($0 \leq k < K-1$; $K = 400$ ($\hat{=}$ 25 ms)). After a windowing and zero padding, the FFT ($N = 512$) is applied. The magnitude spectrum is derived from the positive half of the FFT output $\underline{X}(a,n)$, where n refers to the frequency index. HFA synthesizes spectra $X_s(a,n)$ which are subsequently analyzed by an MFB as used in [1, 9]. The following steps are accomplished within HFA:

- Estimation of the fundamental frequency F_0
- Measuring of the harmonic components of the signal
- Voiced/Unvoiced Detection VUD
- Spectral synthesis dependent on the VUD decision. (Key of the HFA method)

2.3 Detection of the Fundamental Frequency F_0

New advancements are made every year for the detection of fundamental frequency. The influence and robustness against noise was already an issue. However, except Unokis recent investigation [10], studies about the behavior of F_0 detection methods among reverberation conditions are not published so far. While investigating the proposed Method HFA, Unokis study was not released. Thus the authors relied on the results of studies like [11] which shows a very well working performance for the autocorrelation function (ACF) method in the car environment. This study was chosen since in-car noises are mainly distributed at low frequencies where reverberation has its main disturbing region (mentioned above). Hence HFA initially uses the ACF method for the detection of F_0. The positive half of the autocorrelation sequence $r_{xx}(a,\kappa)$ is derived from the actual frame $x(a,k)$. The maximum within the concerning limits of $r_{xx}(a,\kappa)$ indicates the fundamental period duration at the lag κ_0

$$\kappa_0(a) = \arg \max_{\kappa_{0,min}}^{\kappa_{0,max}} r_{xx}(a,\kappa) \tag{1}$$

Suitable for most humans the limits are assigned with

$$\kappa_{0,[max/min]} = \frac{f_s}{F_{0,[min/max]}}; \quad F_{0,[min/max]} = [70/400]Hz \tag{2}$$

The fundamental frequency is finally derived with $F_0(a) = f_s/\kappa_0(a)$. Since the detection of the fundamental frequency is not always reliable a post processing method is applied to enhance the F_0 detection. For post processing the authors choose a simple method which deletes single impulses in the function $F_0(a)$. Strong changes in both directions are detected and substituted with the mean of the previous and the following value of F_0 (ref. Figure 3). Also the authors accomplished experiments using the more sophisticated Viterbi approach as used in [12, 13], which finds the best path between more than one F_0 candidates. However, compared

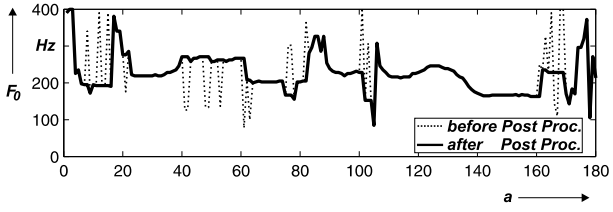

Figure 3: *Detected $F_0(a)$ as a function of the frame index a.*

to the former proposed simple method no gain was found for the reverberant environments using the Viterbi approach. These results correspond also to [11] which found only a neglectable improvement using Viterbi post processing for ACF. For that reason and the higher numerical complexity the usage of the Viterbi approach is omitted. In spite of no existent F_0 for unvoiced frames, firstly a value for $F_0(a)$ is assigned for all frames.

Remark: Within this publication the evaluation of F_0 detection methods is not the key issue. A quantitative assessment of the functioning of the F_0 detection is not given. However, current research of the authors is focused on the study and improvement of F_0 detection methods in reverberant environments.

2.4 Estimation of the Harmonic Components

The harmonic components of $|\underline{X}(a,n)|$ are assumed at harmonic indices $n_{h,i}(a)$ belonging to the i-th multiples of $F_0(a)$

$$\begin{aligned} n_{h,i}(a) &= round\left(\frac{i\cdot F_0(a)\cdot N}{f_s}\right)\Big|_{i=1}^{i_{max}(a)} \\ i_{max}(a) &= round\left(\frac{f_{max}}{F_0(a)}\right) \end{aligned} \tag{3}$$

f_{max} is set to 6000 Hz, but could also be varied. These harmonic indices are used to derive a "spectrum" of harmonics:

$$X_h(a,n) = \begin{cases} |\underline{X}(a,n)| & n \in n_{h,i}(a) \\ 0 & n \notin n_{h,i}(a) \end{cases} \tag{4}$$

2.5 Voiced/Unvoiced Detection

HFA uses a VUD to distinguish between voiced and unvoiced speech frames within utterances. Several F_0 detection approaches include methods for VUD (e.g. the Viterbi post processing [12, 13]). But as already mentioned, studies have taken place only in noisy but not in reverberant environments. Usually VUDs use the value of the peak in the decision function, which represents the strength of F_0. This principle can be applied for several types of decision functions. Here the VUD does not use the strength of this peak. Instead the decision is based on the mean energy $E_h(a)$ of the harmonic components in X_h. They highly exceed the mean of the full band energy for voiced frames, which is more significant than the peak in the decision function.

$$E_h(a) = \frac{1}{i_{max}(a)} \sum_{i=1}^{i_{max}(a)} X_h^2(a, n_{h,i}(a)) \tag{5}$$

The VUD decides for *voiced* if the function $E_h(a)$ exceeds a threshold E_{th}, for *unvoiced* otherwise.

$$E_{th} = \frac{b_{th}}{A} \sum_{a=0}^{A-1} E_h(a) \tag{6}$$

Remark: Also the design of the VUD is not the key issue of this publication. Except the experimentally determination of b_{th} (refer section 3) the performance of the simple VUD approach was not further enhanced, this applies for future work. But Figure 9 already displays that a moderate variation of b_{th} has no deep impact on the RR, which points out that the method holds for moderate changes of the mean energy E_{th} as present in online processing (due to noises or decreasing robustness of the voice activity detector VAD).

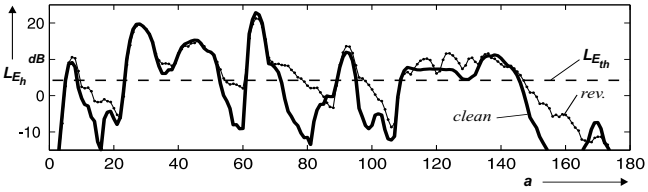

Figure 4: *Plot of logarithmized harmonic energy* $10\lg(E_h(a))$ *for a clean and a reverberant* (T_{60} = *700 ms, SMD = 1 m) signal. The dashed line shows the threshold with* $b_{th} = 0.25$

2.6 Spectral Synthesis

After derivation of the harmonic components and the VUD decision the actual synthesized spectrum $X_s(a,n)$ is derived. The spectral synthesis implements the 3 basic ideas of HFA **(i), (ii), (iii)** in section 2.1. The synthesis is achieved differently for voiced (index v) and unvoiced (index u) frames. In this section the frame index a is omitted, all equations refer to a current frame a. Extensive experiments have been accomplished for the optimization of the subsequently described parameters (fade-in/out windows, exponent). The parameters have been varied and a consistent set was applied for one training and evaluation run (not further described in detail here). An optimal working parameter set is given below.

2.6.1 Treatment of Unvoiced Frames

For frames classified as unvoiced the ideas **(ii)** and **(iii)** of section 2.1 are implemented. High frequency components are left unchanged. Low frequency components are completely deleted and replaced by a noise floor $X_{Fl}(n)$. The noise is chosen instead of zero padding, to achieve a defined low energy level of L_{Fl} = -60 dB representing silence. The logarithm inside of the Mel-Filter-Bank would lead to $-\infty$, if zero padding would be used. L_{Fl} can also be adaptively set (not described here). The upper frequencies are faded in using the first half of a Hamming window (width w_u, start at offset index d_u, refer Figure 5).

$$W(n) = 0.538 - 0.462\cos\left(\frac{2\pi\cdot(n-(d_u+w_u))}{2\cdot w_u}\right) \tag{7}$$

The exact derivation of a synthetic unvoiced spectrum follows

$$X_{s,u}(n) = \begin{cases} X_{Fl}(n) & 0 \le n < d_u \\ |\underline{X}(n)|W(n) & d_u \le n < d_u + w_u \\ |\underline{X}(n)| & d_u + w_u \le n < \frac{N}{2} \end{cases} \tag{8}$$

An optimal parameter set is found with f_{d_u} = 500 Hz, f_{w_u} = 750 Hz, which leads to a suppression of reverberation lower than 1250 Hz.

2.6.2 Treatment of Voiced Frames

For frames classified as voiced the ideas **(i)** and **(iii)** of section 2.1 are implemented. High frequency components are also left unchanged, but the assignment of the parameters d_v and w_v is different compared to the unvoiced case. Contrary low frequency components are synthesized as a synthetic harmonic spectrum $X_{s,h}(n)$ based on the harmonic spectral components $X_h(n)$ achieved in (4). Subsequently the gaps between the spectral lines of $X_h(n)$ in $X_{s,h}(n)$ at spectral indices $n \notin n_{h,i}$ are refilled for energy reconstruction of the

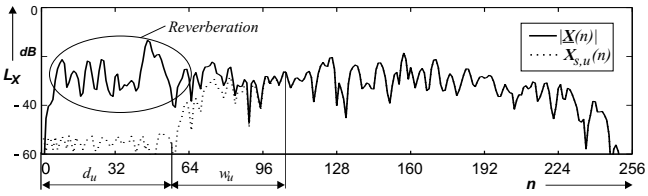

Figure 5: *Logarithmically scaled original $|\underline{X}(n)|$ and associated synthetic unvoiced spectrum $X_{s,v}(n)$ ("s" contaminated with reverberation coming from previous voiced speech).*

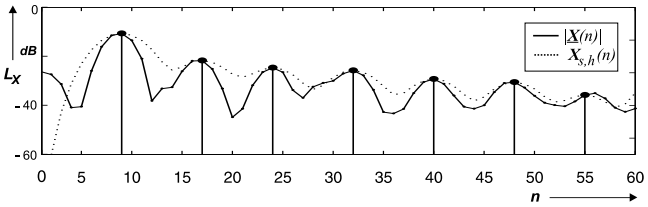

Figure 6: *First 60 spectral components of a logarithmically scaled original voiced spectrum, associated spectrum of harmonics $X_h(n)$ sampled at multiples of F_0 (stems) and associated synthetic harmonic spectrum $X_{s,h}(n)$ with $\gamma = 4$.*

spectrum. The easiest solution for the filling would be the linear interpolation of the harmonic spectral lines. To keep the waveform structure of a voiced speech spectrum the authors chose the addition of weighted von Hann windows as a first approach. At the position of a current index $n_{h,i}$ the von Hann window $W_{n_{h,i}}(n)$ reaches from the previous $(n_{h,i-1})$ to the next $(n_{h,i+1})$ harmonic index (otherwise zero).

$$\begin{aligned} X_{s,h}(n) &= \textstyle\sum_{i=1}^{i_{max}} X_h(n_{h,i}) W_{n_{h,i}}(n) \\ W_{n_{h,i}}(n) &= \left(\tfrac{1}{2} - \tfrac{1}{2}\cos\left(\tfrac{2\pi\cdot(n-n_{h,i})}{n_{h,i+1}-n_{h,i-1}}\right)\right)^{\gamma} \end{aligned} \tag{9}$$

The exponent γ is responsible for the waveform in the synthesized spectrum (refer Figure 6 ($\gamma = 4$)). The synthesis of the final synthetic voiced spectrum $X_{s,v}(n)$ is a combination of the original spectrum $|\underline{X}(n)|$ in the upper frequency regions and the synthetic harmonic spectrum $X_{s,h}(n)$ in the lower frequency regions. The intersection of both regions is handled with a fade-out $W_L(n)$ resp. fade-in $W_H(n)$ window (Hamming) similar to the unvoiced case.

$$\begin{aligned} W_L(n) &= 0.538 - 0.462\cos\left(\tfrac{2\pi\cdot(n-d_v)}{2\cdot w_v}\right) \\ W_H(n) &= 0.538 - 0.462\cos\left(\tfrac{2\pi\cdot(n-(d_v+w_v))}{2\cdot w_v}\right) \end{aligned} \tag{10}$$

The exact derivation of a synthesized voiced spectrum follows

$$X_{s,v}(n) = \begin{cases} X_{s,h}(n) & 0 \le n < d_v \\ X_{s,h}(n)W_L(n) + |\underline{X}(n)|W_H(n) & d_v \le n < d_v + w_v \\ |\underline{X}(n)| & d_v + w_v \le n < \frac{N}{2} \end{cases} \tag{11}$$

An optimal working parameter set is found with $\gamma = 4$; $f_{d_v} =$ 500 Hz; $f_{w_v} =$3500 Hz. That means the method is working well, if frequencies below 4000 Hz are synthesized out of the harmonics.

2.7 Dealing with VUD Classification Errors

For a current frame the VUD classification result calls for one of two different analysis methods, which would generate distant feature vectors in the feature space for the same frame. HFA has to consider an uncertain behavior of the VUD, which even increases in presence of disturbances as reverberation. These errors appear in the training and in the recognition phase. Therefore during the training two distant vector clusters (a "cleaned" and a "disturbed" one; or in other words one for the right and one for the wrong VUD decision) are formed in the feature space belonging to the same phoneme class. Vector quantisation and the use of multiple gaussians for the same phoneme model cover this effect. The wrong VUD decisions have the following effects:

a) ***Voiced frame classified as unvoiced:*** Low frequency components of voiced frames are hardly suppressed. In opposite to unvoiced frames this strong impact deletes important information from the voiced signal. Remaining are frequency components above the fade-in frequencies (500 ... 1250 Hz , experimental optimization). Hence for most vowels the first formant information will be deleted but the information of the second one (and higher) remain. Hence the distant vector cluster will still contain reduced information to distinguish between voiced phonemes. This is one reason why the parameter optimization resulted in much lower fade-in frequencies than 2500 Hz as in idea **(iii)**.

b) ***Unvoiced frame classified as voiced:*** The F_0 detection suggests a value between $F_{0,min} \ldots F_{0,max}$. This may be an imaginary F_0 or a smeared F_0 from a previous voiced section (refer Figure 1 (right)). In both cases the harmonic synthesis generates a disturbed unvoiced harmonic synthesized spectrum. Due to reverberant training the "disturbed" unvoiced clusters have distributions with higher variances coming from different previous voiced sections in the training material. For this VUD decision case HFA does not work better than CFA with MCT.

Compared to clean training the number of wrong VUD classifications even increases at reverberant training. This is in fact an advancing effect, since a better modeling of the "disturbed" clusters of the phonemes is performed.

3. EXPERIMENTAL RESULTS

This work is based on the previous work [1], hence the same UASR recognizer [9] (30 channel MFB, 44 monophone HMMs with GMM) and evaluation setup is used (1020 command phrases as subset of the APOLLO corpus [14] containing 17 classes). The reverberation conditions are achieved by convolving the evaluation data with real measured RIRs (T_{60} = [0, 200, 400, 700, 1000, 2000] ms; $SMDs$ = [100, 300] cm). In addition to the clean training a MCT is ap-

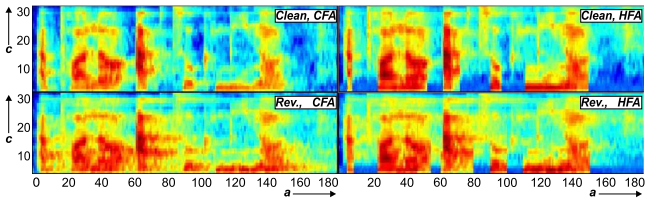

Figure 7: *Feature matrix of a clean and a reverberant (T_{60} = 700 ms ; SMD = 100 cm) utterance. HFA reduces the pattern mismatch for varying conditions compared to CFA.*

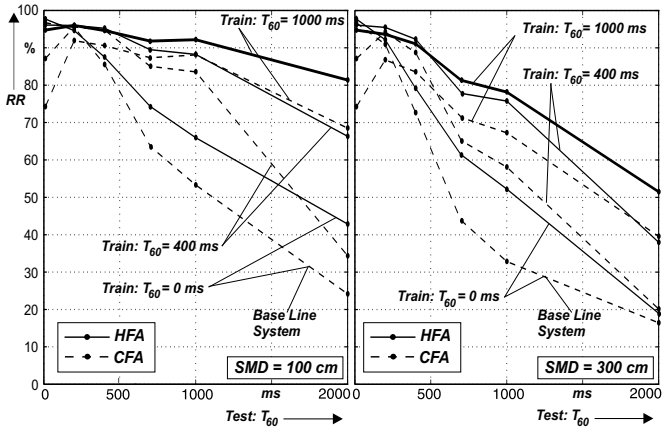

Figure 8: *Evaluation results for SMD = 100 cm (left) and 300 cm (right). RR performances of CFA and HFA at varied training conditions (T_{60} = [0, 400, 1000] ms)*

plied, which means also the training data is reverberated to gain different training conditions (T_{60} = [0, 400, 1000] ms; SMD = 100 cm). In Figure 8 the main evaluation results are displayed. Here HFA competes with the CFA method. As in related publications, it can be seen that the baseline system (CFA, clean training (T_{60} =0 ms)) performs very poor with increasing reverberation times and increasing SMDs (difference left vs. right). For the clean training the HFA method already increases the performance compared to CFA. As mentioned in section 2.7 the reverberant training increases the performance of HFA by better modeling of the "disturbed" vector cluster and including the reverberant condition into the "clean" vector cluster. Reverberant training also increases the performance of the CFA, but only around the special training condition - e.g. clean evaluation data performs worse for reverberant training. Also varying *SMD*s (left vs. right) have a deep impact on the RR of CFA using MCT. Contrary HFA with MCT increases the RR for a wide range of reverberation conditions, also for varying *SMD*s. The best performance could be achieved for the training condition T_{60} = 1000 ms; SMD = 100 cm. But also training at T_{60} = 400 ms performs almost sufficient, taking into account that tests conditions at T_{60} = 2000 ms are already extreme environments (stair case). Finally a number of experiments for optimization of the VUD threshold parameter b_{th} are conducted. For several settings of b_{th} = [0.1 0.2 0.5 1.0 1.7] the model was trained at the well working condition T_{60} = 1000 ms; SMD = 100 cm. Figure 9 shows the results for the test of the model trained with b_{th} = 1.0, while b_{th} was changed during test to simulate online conditions. For most observed T_{60} varying b_{th} has only limited effect on the RR (optimum at $b_{th} = 1$ ($\hat{=}$ 0 dB)). This emphasizes also the explanations given in 2.7.

4. CONCLUSION AND FUTURE WORK

The authors have proved that ASR under (changing and even strong) reverberation conditions is possible using the introduced method HFA. HFA achieves a higher robustness than the CFA by emphasizing the direct/clean speech components and suppression of the reverberation components. The mismatch between the feature patterns at different reverberation conditions is decreased using HFA in comparison to CFA (Figure 7). This yields to stable and matching test and

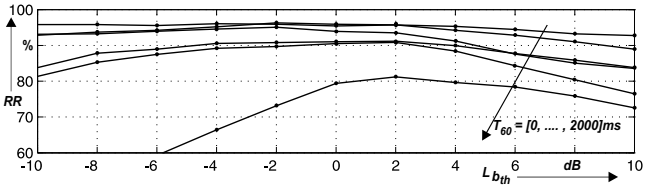

Figure 9: *Optimization of b_{th}: RR of the 6 test conditions (T_{60} = [0, 200, 400, 700, 1000, 2000] ms; SMDs = 100 cm) for different values of b_{th} ($L_{b_{th}}$ in dB). Training condition T_{60} = 1000 ms; SMDs = 100 cm; b_{th} = 1 ($\hat{=}$ 0 dB)*

training conditions. Apart from the high performance of the ASR system the advantages of HFA compared to state of the art blind dereverberation algorithms are real time processing (no adaptation time is needed) and higher robustness against moderate speaker movements. Current work of the authors is a detailed assessment of F_0 detection methods among reverberant environments, an improvement of the VUD and the assessment of HFA in comparison with the MTF based method [3]. Tests with MFCCs instead of MFB and the combination with noise reduction methods is also an issue for the future.

REFERENCES

[1] Petrick, R., Lohde, K., Wolff, M., Hoffmann, R.: "The harming part of room acoustics for automatic speech recognition", Proc. of INTERSPEECH 2007, Antwerp, 2007.

[2] Kinoshita, K., Delcroix, M., Nakatani, T., Miyoshi, M.: "Multi-step linear prediction based speech dereverberation in noisy reverberant environment", Proc. of INTERSPEECH 2007, Antwerp, 2007.

[3] Unoki, M., Furakawa, M., Sakata, K., Akagi, M.: "An improved method based on the MTF concept for restoring the power envelop from reverberant signal", Acoust. Sci. & Tech., vol. 25, pp. 232-242, 2004.

[4] Gillespie, B. W., Malvar, H. S., Florencio, D. A.: "Speech dereverberation via maximum-kurtosis subband adaptive filtering", Proc. of ICASSP 2001, Salt Lake City, Utah, USA, May 2001.

[5] Nakatani, T., and Miyoshi, M.: "Blind dereverberation of single channel speech signal based on harmonic structure", Proc. ICASSP 2003, vol. 1, pp. 92-95, Hong Kong, April 2003.

[6] Leggetter, C. J., Woodland, P. C.: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models", Computer Speech and Language, vol. 9, pp. 171-185, 1995.

[7] Sehr, A., Zeller, M., Kellermann, W.: "Hands-free speech recognition using a reverberation model in the feature domain", Proc. European Signal Processing Conference (EUSIPCO), Florence, Italy, Sep. 2006.

[8] Nakatani, T., Kinoshita, K., and Miyoshi, M., "Harmonicity based blind dereverberation for single-channel speech signals," IEEE Trans. Audio, Speech, and Language Processing, vol.15, no.1, pp.80-95, 2007.

[9] Hoffmann, R.; Eichner, M.; 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. Berlin etc.: Springer-Verlag 2007, LNAI 4775, 200-218.

[10] Unoki, M., Hosorogiya, T.: "Estimation of fundamental frequency of reverberant speech by utilizing complex cepstrum analysis", Journal of Signal Processing, vol. 12, No. 1, pp. 31 - 44, Jan. 2008.

[11] Quast, H., Schreiner, O., Schroeder, M. R.: "Robust pitch tracking in the car environment", Proc. of ICASSP 2002, Orlando, USA, 2002.

[12] Luengo, I., Saratxaga, I., Navas, E., Hernaez, I., Sanchez, J., Sainz, I.: "Evaluation of pitch detection algorithms under real conditions", Proc. of ICASSP 2007, Honolulu, Hawaii, USA, April 2007.

[13] Boersma, P.: "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", IFA Proceedings vol. 17, University of Amsterdam, Institute of Phonetic Sciences, pp. 97-110, Amsterdam, Netherlands, 1993.

[14] Maase, J., Hirschfeld, D., Koloska, U., Westfeld,T., Helbig, J.: "Towards an evaluation standard for speech control concepts in real-world scenarios", Proc. of EUROSPEECH 2003, Geneva, 2003.