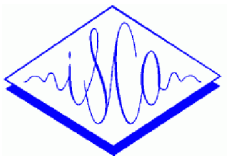

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Bayesian Feature Enhancement for ASR of Noisy Reverberant Real-World Data

Alexander Krueger[1], *Oliver Walter*[2], *Volker Leutnant*[2] *and Reinhold Haeb-Umbach*[2]

[1]Research & Innovation, Technicolor, 30625 Hannover, Germany
[2]Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany
alexander.krueger@technicolor.com, {walter,leutnant,haeb}@nt.uni-paderborn.de

10.21437/Interspeech.2012-246

Abstract

In this contribution we investigate the effectiveness of BAYESIAN feature enhancement (BFE) on a medium-sized recognition task containing real-world recordings of noisy reverberant speech. BFE employs a very coarse model of the acoustic impulse response (AIR) from the source to the microphone, which has been shown to be effective if the speech to be recognized has been generated by artificially convolving non-reverberant speech with a constant AIR. Here we demonstrate that the model is also appropriate to be used in feature enhancement of true recordings of noisy reverberant speech. On the Multi-Channel Wall Street Journal Audio Visual corpus (MC-WSJ-AV) the word error rate is cut in half to 41.9% compared to the ETSI Standard Front-End using as input the signal of a single distant microphone with a single recognition pass.

Index Terms: bayesian feature enhancement, dereverberation, denoising

1. Introduction

The automatic recognition of reverberant speech is currently a "hot" topic, as is evidenced by the tremendous increase of publications on this issue in journals and at international conferences in the last years [1]. However, most of the publications consider small recognition tasks and artificially reverberated data.

While this constraint is valid for an initial treatment of the topic, a comprehensive and meaningful investigation must address the issue of how the proposed approaches scale to larger tasks and how they behave on true recordings from a reverberant environment, where the AIR is no longer constant and where additive noise is typically present, in addition to reverberation.

Approaches to robust speech recognition can be classified into front-end and back-end methods. Back-end methods modify the acoustic models or the decoder to account for the effect of reverberation. It is likely that the computational complexity of these techniques rises as the size of the acoustic model and the recognition task increases. The complexity of front-end techniques, on the other hand, tends to be independent of the size of the recognition task. In front-end techniques, reverberation can be either addressed at the signal or at the feature level. The computation of MEL frequency cepstral coefficients (MFCCs) or similar features incorporates both a decimation in time (the framing) and in frequency (by the MEL filter bank). As a consequence, to describe the impact of reverberation, the AIR need not be known but rather a representation of it in the feature domain, which may be easier to obtain.

For these reasons we decided in prior work to treat reverberation in the feature domain. In [2] we have developed a BAYESIAN feature enhancement approach for reverberant speech recognition and obtained good recognition results on the AURORA5 database, a connected digits recognition task, where the utterances of the database were created by convolving the corresponding clean utterances with artificial AIRs [3].

However, so far we have never verified our claim, that with feature based methods high recognition accuracy can also be obtained on real recordings of noisy reverberant speech. This paper is going to fill this gap and investigates the performance of BFE on the MC-WSJ-AV corpus, a 5000-word recognition task, recorded in a meeting room environment [4].

The paper is organized as follows. In the next section we describe the impact of noise and reverberation on the logarithmic MEL power spectral coefficients (LMPSCs), an intermediate representation during the computation of the MFCCs, followed by a brief introduction to the BFE approach for the ASR of noisy reverberant speech in Sec. 3. Next, we explain our feature domain model of the AIR in Sec. 4. In the experimental section we first describe the MC-WSJ-AV corpus followed by a number of recognition experiments, which demonstrate the effectiveness of BFE on a medium-sized vocabulary task of true recordings from a reverberant environment.

2. Signal Model

A system theoretic model of reverberation in the time domain is the convolution of the source (clean) speech signal $x(l)$ with the acoustic impulse response $h(l)$ from the source to the sensor. As a distant microphone will also capture additive noise $n(l)$, the signal at the microphone $y(l)$ can be expressed as

$$y(l) = \sum_{p=0}^{L_h-1} h(p)x(l-p) + n(l), \qquad (1)$$

where L_h denotes the length of the AIR. When the short-time discrete Fourier transform (STDFT) is applied to (1), the following relationship results among the STDFTs $\tilde{Y}(m,f)$, $\tilde{X}(m,f)$ and $\tilde{N}(m,f)$ of the noisy reverberant speech, the clean speech and the noise signal, respectively:

$$\tilde{Y}(m,f) \approx \sum_{m'=0}^{L_H} \tilde{X}(m-m',f)H_{m',f} + \tilde{N}(m,f), \qquad (2)$$

where the number of summands in the convolution is given by

$$L_H = \left\lfloor \frac{L_h + L_w - 2}{B} \right\rfloor, \qquad (3)$$

with L_w denoting the window length, B being the frame shift and $\lfloor\cdot\rfloor$ denoting the floor function. In (2), m indicates the frame index, while f is the frequency bin index. $H_{m,f}$ is a frequency domain representation of the windowed AIR. Note that the convolution in (1) results in a convolution in the STDFT

domain, however now w.r.t. the frame index m.

After the power spectrum computation, application of the MEL filter bank and transformation into the logarithmic domain, the relationship between nonreverberant, noise, and noisy reverberant features becomes highly nonlinear [2]:

$$\begin{aligned}\mathbf{y}_m &= \ln\left(\sum_{m'=0}^{L_H} e^{\mathbf{x}_{m-m'}+\overline{\mathbf{h}}_{m'}} + e^{\mathbf{n}_m}\right) + \mathbf{v}_m \\ &= f\left(\mathbf{x}_{m-L_H:m}, \overline{\mathbf{h}}_{0:L_H}, \mathbf{n}_m\right) + \mathbf{v}_m. \end{aligned} \quad (4)$$

Here, $\mathbf{y}_m$, $\mathbf{x}_m$ and $\mathbf{n}_m$ are the LMPSC feature vectors of the noisy reverberant speech, the nonreverberant speech and of the noise signal, respectively, at frame m. $\overline{\mathbf{h}}_m$ is an approximate representation of the AIR in the logarithmic MEL spectral domain. Note that the nonlinearity $f(\cdot)$ has to be understood to be applied element-wise to the vectors and that

$$\mathbf{x}_{m-L_H:m} := \mathbf{x}_{m-L_H}, \ldots, \mathbf{x}_m \quad (5)$$

$$\overline{\mathbf{h}}_{0:L_H} := \overline{\mathbf{h}}_0, ..., \overline{\mathbf{h}}_{L_H} \quad (6)$$

denote the sequence of LMPSC feature vectors of clean speech and the components of the representation of the AIR, respectively.

The vector $\mathbf{v}_m$ captures all errors resulting from the various approximations that had to be introduced to arrive at (4). In [2, 5] we have proposed to model $\mathbf{v}_m$ as a realization of a white GAUSSIAN process whose mean vector and covariance matrix can be estimated by comparing the observed LMPSC feature vectors with those predicted by the model.

3. BAYESIAN Feature Enhancement

In this section, we will briefly explain the idea of BAYESIAN feature enhancement. For a detailed treatment the reader is referred to [2, 6].

Assuming we are given a sequence of M jointly reverberant and noisy feature vectors

$$\mathbf{y}_{1:M} := \mathbf{y}_1, \ldots, \mathbf{y}_M, \quad (7)$$

the goal is to estimate the sequence of the corresponding clean feature vectors $\mathbf{x}_{1:M}$. We assume that we can afford a latency of $L_\mathrm{C} - 1$ frames to be able to exploit the knowledge of the sequence $\mathbf{y}_{1:m+L_\mathrm{C}-1}$ for the estimation of $\mathbf{x}_m$. The estimation can greatly benefit from the use of future observations, as will be seen in the experimental results.

The estimation of the clean feature vectors is formulated as a trajectory tracking problem, where all involved kinds of acoustic feature vector trajectories are assumed to be realizations of vector valued stochastic processes. For the estimation, we introduce the state vector

$$\mathbf{z}_m := \left(\left(\boldsymbol{\chi}_m\right)^T, \left(\mathbf{n}_m\right)^T\right)^T, \quad (8)$$

where $\boldsymbol{\chi}_m$ is a vector consisting of the clean LMPSC vector at time instant m and $L_\mathrm{C} - 1$ previous clean LMPSC vectors according to

$$\boldsymbol{\chi}_m := \left(\left(\mathbf{x}_m\right)^T, \ldots, \left(\mathbf{x}_{m-L_\mathrm{C}+1}\right)^T\right)^T. \quad (9)$$

The goal of BAYESIAN feature enhancement is to obtain the posterior probability density function (PDF) $p\left(\mathbf{z}_m | \mathbf{y}_{1:m}\right)$. From this we obtain the minimum mean square error (MMSE) estimate $\hat{\mathbf{x}}_{m-L_\mathrm{C}+1} = \mathrm{E}\left[\mathbf{x}_{m-L_\mathrm{C}+1} | \mathbf{y}_{1:m}\right]$, with $\mathrm{E}\left[\cdot\right]$ denoting the expected value, to which a Discrete Cosine Transform (DCT) is applied before being forwarded to the decoder as cleaned-up MFCC feature vector.

The posterior PDF $p\left(\mathbf{z}_m | \mathbf{y}_{1:m}\right)$ may be theoretically computed recursively by carrying out BAYESIAN inference, i.e., performing in turn the so-called prediction

$$\begin{aligned} &p\left(\mathbf{z}_m | \mathbf{y}_{1:m-1}\right) \\ &= \int p\left(\mathbf{z}_m | \mathbf{z}_{m-1}, \mathbf{y}_{1:m-1}\right) p\left(\mathbf{z}_{m-1} | \mathbf{y}_{1:m-1}\right) d\mathbf{z}_{m-1} \end{aligned} \quad (10)$$

and the update step

$$p\left(\mathbf{z}_m | \mathbf{y}_{1:m}\right) \propto p\left(\mathbf{y}_m | \mathbf{z}_m, \mathbf{y}_{1:m-1}\right) p\left(\mathbf{z}_m | \mathbf{y}_{1:m-1}\right). \quad (11)$$

While the prediction step requires the knowledge of the predictive PDF $p\left(\mathbf{z}_m | \mathbf{z}_{m-1}, \mathbf{y}_{1:m-1}\right)$, the update step asks for the knowledge of the observation PDF $p\left(\mathbf{y}_m | \mathbf{z}_m, \mathbf{y}_{1:m-1}\right)$.

For the predictive PDF we employ a switching linear dynamic model (SLDM) to capture the dynamics of speech features, and a GAUSSIAN PDF as a priori model of noise. A SLDM consists of I interacting sub-models, where the i-th submodel, indicated by the regime variable ζ_m, is described by

$$p\left(\mathbf{x}_m | \mathbf{x}_{m-1}, \zeta_m = i\right) \quad (12)$$

$$\approx \begin{cases} \mathcal{N}\left(\mathbf{x}_m; \boldsymbol{\mu}_{\mathbf{x},i}, \boldsymbol{\Sigma}_{\mathbf{x},i}\right) & \text{for } m = 1 \\ \mathcal{N}\left(\mathbf{x}_m; \mathbf{A}_i \mathbf{x}_{m-1} + \mathbf{b}_i, \mathbf{V}_i\right) & \text{for } m > 1. \end{cases} \quad (13)$$

In (13), $\mathcal{N}\left(\cdot; \boldsymbol{\mu}_{\mathbf{x},i}, \boldsymbol{\Sigma}_{\mathbf{x},i}\right)$ denotes a GAUSSIAN PDF with mean vector $\boldsymbol{\mu}_{\mathbf{x},i}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{x},i}$. Further, $\mathbf{A}_i$, $\mathbf{b}_i$ and $\mathbf{V}_i$ denote the state transition matrix, the bias compensation vector and the linear prediction error covariance matrix, respectively.

The observation PDF $p\left(\mathbf{y}_m | \mathbf{z}_m, \mathbf{y}_{1:m-1}\right)$ is derived from the relation (4). This requires $\overline{\mathbf{h}}_m$, a representation of the AIR in the LMPSC domain, which will be described in the next section.

Note that both the a priori model of speech and the observation mapping $f(\cdot)$ are nonlinear. As a consequence, only an approximate recursive BAYESIAN inference of the posterior is computationally feasible. We employed the Interacting Multiple Model algorithm with a bank of iterated extended KALMAN filters for this purpose [7].

4. Feature Domain Model of AIR

We model the AIR as a realization of a zero-mean white Gaussian process $\xi(l)$ of finite length and unit power with an exponentially decaying envelope, a model proposed in [8]:

$$h(l) = \sigma_h \cdot u(l) \cdot \xi(l) \cdot e^{-\frac{l}{\tau_h}}. \quad (14)$$

Here, $u(l)$ is an indicator function constraining the length of the AIR to L_h: $u(l) = 1$ for $0 \leq l < L_h$ and $u(l) = 0$ else. This model has two parameters. The first, σ_h, controls the energy E_h of the AIR, since

$$E_h := \mathrm{E}\left[\sum_{l=0}^{L_h-1} h^2(l)\right] = \sigma_h^2 \frac{1 - e^{-2L_h/\tau_h}}{1 - e^{-2/\tau_h}}. \quad (15)$$

The second parameter is related to the room reverberation time T_{60} through

$$\tau_h = \frac{T_{60}}{3\ln(10) \cdot T_\mathrm{A}} \quad (16)$$

with T_A denoting the sampling interval. Based on (14) a reasonable length L_h of the AIR model may be determined in dependence on τ_h by

$$L_h = L_h(\tau_h) = \left\lceil -\frac{\tau_h}{2} \ln(\epsilon_h) \right\rceil, \quad (17)$$

which is obtained by minimizing the AIR length under the constraint that the relative energy of the neglected part of the AIR

is smaller than some prespecified ϵ_h. In (17) $\lceil \cdot \rceil$ denotes the ceiling function.

The advantage of using this model is that is has only two parameters, which can be estimated more easily than the complete AIR. An estimate of τ_h can be computed from an estimate of the reverberation time by (16), an estimate of σ_h can be obtained from an estimate of E_h and (15).

On databases containing artificially reverberated data, the estimation of E_h is usually superfluous since the AIR is often normalized to unit energy. However, this does in general not hold for true recordings of reverberant speech. In that case, the condition $E_h = 1$ can be assured to approximately hold after an appropriate normalization of the test data, which can be derived as follows: First, it is assumed that the average power of reverberant speech, σ_s^2, is approximately related to that of non-reverberant speech, σ_x^2, through $\sigma_s^2 \approx E_h \sigma_x^2$. Second, it is assumed that the average power of the non-reverberant speech of the test data, $\sigma_{x,\mathrm{TEST}}^2$, is equal to that of the training data, $\sigma_{x,\mathrm{TRAIN}}^2$. To assure $E_h = 1$, the noisy reverberant test data is multiplied by $\sqrt{\sigma_{s,\mathrm{TEST}}^2 / \sigma_{x,\mathrm{TRAIN}}^2}$ before feature enhancement, where the average power of reverberant speech of the test data, $\sigma_{s,\mathrm{TEST}}^2$, can be computed from the average power of the noisy reverberant speech in the test data, $\sigma_{y,\mathrm{TEST}}^2$, and the average power of the noise in the test data, $\sigma_{n,\mathrm{TEST}}^2$, by $\sigma_{s,\mathrm{TEST}}^2 \approx \sigma_{y,\mathrm{TEST}}^2 - \sigma_{n,\mathrm{TEST}}^2$ and by using a voice activity detector.

In Monte Carlo simulations it was observed that the distributions of logarithmic MEL power spectral representation of the AIR $\overline{\mathbf{h}}_{0:L_H}$, Eq. (6), under the stochastic AIR model (14) can be well approximated by GAUSSIANs. As we have previously done in [2], we replace the usually unknown representation $\bar{\mathbf{h}}_{0:L_H}$ in the observation mapping (4) by the means $\boldsymbol{\mu}_{\bar{\mathbf{h}}_0}, \ldots, \boldsymbol{\mu}_{\bar{\mathbf{h}}_{L_H}}$ of these GAUSSIANs.

5. Experimental Results

5.1. The MC-WSJ-AV Corpus

The Multi-Channel Wall Street Journal Audio Visual corpus (MC-WSJ-AV) is a collection of read Wall Street Journal (WSJ) sentences taken from the development and evaluation sets of the WSJCAM0 database, recorded in a number of instrumented meeting rooms constructed within the framework of the European AMI (Augmented Multi-Party Interaction) project [4]. Sentences are read according to three different scenarios, where the experiments reported here have been conducted on the "single stationary speaker" subset. For this condition the speakers read sentences from six positions within the meeting room — four seated around a table, one standing at the whiteboard, and one standing at the presentation screen. Data have been recorded simultaneously by a headset microphone, a lapel microphone and two 8-element circular microphone arrays positioned on the table. The test set used in the experiments reported here are the EVAL1 sentences recorded at the University of Edinburgh, consisting of 189 sentences, totaling 3093 words and having a total length of 21 minutes. The sampling rate is $16\,\mathrm{kHz}$.

The signal-to-noise ratio at the input of a microphone of the circular array is on average about $15 - 20\,\mathrm{dB}$ and the room reverberation time was estimated to $700\,\mathrm{ms}$ [4] and $380\,\mathrm{ms}$ [9].

Several authors have published recognition results on this database [4, 9]. They mostly experimented with microphone array techniques to reduce the impact of reverberation.

5.2. Baseline Experiments

Acoustic model training has been carried out with the WSJ-CAM0 training set using the HTK system. The acoustic model consisted of approximately 9000 tied-state triphones with three emitting states per triphone and 10 mixture components per state. The 39-element feature vector comprised of 13 MFCCs, including the 0th cepstral coefficient, with their first and second-order derivatives, was computed according to the ETSI Standard Front-End [10], with the modification that power instead of magnitude spectrum was employed. The language model is the standard MIT-Lincoln Labs 5k Wall Street Journal bigram language model.

Unlike many artificially reverberated databases, here the energy of the AIR is not normalized to unity. As a result, we normalized the test data as described in the previous section.

Table 1 displays baseline recognition results using standard feature extraction techniques: The ETSI Standard Front-End (SFE) [10], followed by cepstral mean normalization (CMN) on the full 39-dimensional feature vector, and the ETSI Advanced Front-End (AFE) [11]. The results obtained on the headset, lapel and single distant microphone (SDM, we used microphone 1 of array 1 for our tests) are comparable to those published in [4]. Note that a second recognition pass employing unsupervised channel adaptation by a single Constrained Maximum Likelihood Linear Regression (CMLLR) transformation matrix [12], trained on the transcription provided by the first recognition pass, did only deliver a relatively small improvement.

Channel	Approach	WER [%] no adapt.	WER [%] CMLLR
Headset	SFE + CMN	15.3	14.7
Lapel	SFE + CMN	23.5	18.9
SDM	SFE + CMN	82.5	73.4
SDM	AFE	74.4	66.2

Table 1: Baseline recognition results on MC-WSJ-AV

5.3. Bayesian Feature Enhancement

In this section we present the performance of the BFE approach for ASR of noisy reverberant speech, using the SDM data, in terms of word error rates achieved. The processing steps are as follows: First, 23-component LMPSC feature vectors are computed according to the slightly modified ETSI SFE. Then BFE is applied and the enhanced features are transformed by DCT to 13 static MFCCs. Next, velocity and acceleration features are appended and CMN is applied to the resulting 39-dimensional feature vector, which is finally forwarded to the decoder.

In a first set of experiments the settings of the parameters of the AIR model of Sec. 4 are determined. For these experiments we used a reduced complexity system with only $I = 4$ dynamic models within the SLDM, the a priori model of speech. Further, the latency parameter was set to $L_C = 4$.

After scaling the test data to unit AIR energy as described in the previous section, the first parameter of the AIR model, the energy parameter σ_h, can be readily computed from (15).

The decay constant has to be set according to the room reverberation time as shown in (17). The AIR model is, however, only a crude approximation to the true AIR. Therefore, the choice of T_{60} optimal with respect to the recognition rate may be different from the true reverberation time of the room

the data had been recorded in. We therefore ran a set of recognition experiments to determine the dependence of the word error rate on the assumed room reverberation time, see Table 2.

The results reveal that the word error rate (WER) is not very sensitive to the choice of room reverberation time: while the lowest WER is obtained for $\hat{T}_{60} = 600\,\mathrm{ms}$, the error rate increases only little if $\hat{T}_{60}$ is reduced or increased by $50\,\mathrm{ms}$. An accuracy within $\pm 50\,\mathrm{ms}$ is what can be expected by state-of-the-art algorithms for the blind estimation of the room reverberation time. From the obtained results we concluded to set $\hat{T}_{60} = 600\,\mathrm{ms}$ for the following experiments. The value is needed in (17) for the determination of L_h to be used for the computation of the logarithmic MEL power spectral representation of the AIR.

$\hat{T}_{60}$ [ms]	WER [%]
400	52.6
450	48.8
500	46.1
550	45.5
600	45.3
650	46.6
700	49.4

Table 2: Word error rates achieved with BFE of SDM input as a function of assumed room reverberation time $\hat{T}_{60}$

Having set the parameters of the AIR model we are going to present the results of a full-fledged system, incorporating a SLDM with $I = 14$ dynamic models. Note that the computational complexity of BFE is roughly proportional to the number of dynamic models I.

Latency L_C	WER [%] no adapt.	WER [%] CMLLR
1	74.3	65.9
2	51.4	46.8
3	44.6	41.9
4	43.0	40.5
5	41.9	40.5
6	42.2	41.1

Table 3: Word error rates achieved with BFE of SDM input as a function of latency parameter L_C

Table 3 presents word error rates as a function of the latency parameter L_C. According to Sec. 3, L_C is the number of successive speech feature vectors present in the state variable $\boldsymbol{\chi}_m$ of the KALMAN filters. Note that the computational complexity of the BFE is roughly proportional to L_C. The lowest WER is obtained with $L_C = 5$, while the error rate increases when decreasing L_C. This shows the benefit of using future observations for the enhancement. The obtained WER of 41.9% is to our knowledge the lowest WER achieved on this database with a SDM input and one-pass recognition. Note, that no speaker adaptation has been carried out so far.

The real time factor for the feature enhancement in this setup was 1.4 on an Intel Core i7/$3.20\,\mathrm{GHz}$ CPU using 3 cores. Note that the computational complexity is independent of the acoustic model and the vocabulary.

Using the recognized word sequence to estimate a CMLLR matrix to be used for channel adaptation in a second recognition pass improved the word error rate by only a few percent. The improvement is probably due to the fact that the poor modeling of the direct path and early reflections in the model (14) is in part compensated by the CMLLR adaptation.

6. Conclusions

In this contribution we have presented recognition results for automatic recognition of single distant microphone noisy reverberant speech using our previously proposed BAYESIAN feature enhancement approach. For the first time, recognition experiments have been conducted on true recordings of noisy reverberant speech rather than on artificially reverberated data. On the MC-WSJ-AV task, a 5000-word recognition task, the word error rate could be roughly cut in half. This is comparable to our earlier results obtained on an artificially reverberated version of the WSJ 5k test set [5]. There, also a factor of two in error rate reduction could be achieved, however at lower absolute values of the error rate.

We have further shown that BFE can be combined favorably with CMLLR: channel adaptation after feature enhancement delivers a further small improvement in word error rate.

7. References

[1] T. Nakatani, W. Kellermann, P. Naylor, M. Miyoshi, and B. H. Juang, "Introduction to the special issue on processing reverberant speech: Methodologies and applications," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1673 –1675, sept. 2010.

[2] A. Krueger and R. Haeb-Umbach, "Model-based feature enhancement for reverberant speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1692–1707, 2010.

[3] H. Hirsch, "Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments," Niederrhein University of Applied Sciences, Tech. Rep., 2007.

[4] M. Lincoln, "The multi-channel wall street journal audio-visual corpus (MC-WSJ-AV): Specification and initial experiments," in *in IEEE Autom. Speech Recognition Understanding Workshop (ASRU)*, 2005, pp. 357–362.

[5] A. Krueger and R. Haeb-Umbach, "A model-based approach to joint compensation of noise and reverberation for speech recognition," in *Robust Speech Recognition of Uncertain or Missing Data*, R. Haeb-Umbach and D. Kolossa, Eds. Springer, 2011.

[6] A. Krüger, "Modellbasierte Merkmalsverbesserung zur robusten automatischen Spracherkennung in Gegenwart von Nachhall und Hintergrundstörungen," Ph.D. dissertation, University of Paderborn, Germany, Dec. 2011.

[7] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*. Wiley, New York, 2001.

[8] J. Polack, "La transmission de l' énergie sonore dans les salles," Dissertation, Université du Maine, 1988.

[9] K. Kumatani, J. McDonough, D. Klakow, P. Garner, and W. Li, "Adaptive beamforming with a maximum negentropy criterion," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, may 2008, pp. 180 –183.

[10] ETSI, "ETSI standard document, speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms, etsi es 201 108 v1.1.3," ETSI ES 201 108 V1.1.3, ETSI ES 201 108, Tech. Rep., 2003.

[11] ——, "ETSI standard document, speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms, v1.1.5," ETSI ES 202 050, Tech. Rep., 2007.

[12] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.