A HIGHLY EFFICIENT OPTIMIZATION SCHEME FOR REMOS-BASED DISTANT-TALKING SPEECH RECOGNITION

Roland Maas[1], Armin Sehr[1], Martin Gugat[2], and Walter Kellermann[1]

[1]Multimedia Communications and Signal Processing,
{maas,sehr,wk}@lnt.de
[2]Applied Mathematics II,
gugat@am.uni-erlangen.de
University of Erlangen-Nuremberg, 91058 Erlangen, Germany

ABSTRACT

A highly efficient decoding algorithm for the REMOS (REverberation MOdeling for Speech recognition) concept for distant-talking speech recognition as proposed in [1] is suggested to reduce the computational complexity by about two orders of magnitude and thereby allowing for first real-time implementations. REMOS is based on a combined acoustic model consisting of a conventional hidden Markov model (HMM), modeling the clean speech, and a reverberation model. During recognition, the most likely clean-speech and reverberant contributions are estimated by solving an inner optimization problem for logarithmic melspectral (logmelspec) features. In this paper, two approximation techniques for the inner optimization problem are derived. Connected digit recognition experiments confirm that the computational complexity is significantly reduced. Ensuring that the global optima of the inner optimization problem are found, the decoding algorithm based on the proposed approximations even increases the recognition accuracy relative to interior point optimization techniques.

1. INTRODUCTION

Modern state-of-art speech recognition systems already achieve good performance. The main restriction however is that the user needs to get close to a microphone to communicate conveniently with the system. To further increase the user comfort, it would therefore be desirable to install distant-talking microphones at fixed points, e.g., at the automatic speech recognition (ASR) device itself, so that the users can move freely while communicating with the system.

Since the distance between speaker and microphone in such a hands-free scenario usually is in the range of one to several meters, there are two kinds of distortions that hamper ASR. Besides the desired signal, the microphone picks up reverberation of the desired signal and unwanted additive signals, like background noise or interfering speakers. While significant progress has already been reported within the last decade regarding the robustness of ASR to additive distortions, ASR for highly reverberant environments has only recently attracted increasing attention [2].

While the usual model adaptation techniques, which have been successfully applied in noisy environments, are not suitable for reverberation significantly exceeding the frame length of the recognizer, [3] suggests a model adaptation approach designed particularly for long reverberation. A very promising approach for obtaining an ASR system for reverberant environments is to train a conventional HMM-based recognizer using reverberant data as suggested in [4] and [5]. However, both model adaptation techniques and reverberant training suffer from the conditional indepence assumption underlying any HMM-based system, namely that the current output vector depends only on the current state. This assumption prevents conventional HMMs from appropriately modeling reverberation which increases the correlation between neighboring frames.

In [1], the REMOS concept is proposed to overcome the limitation of the conditional independence assumption of HMMs. The approach is based on combining a network of clean speech HMMs and a reverberation model (RVM). In the decoding phase, the most likely contribution of the HMM output and the reverberation model output to the current reverberant observation is found by an inner optimization operation for logmelspec features. It has been shown in [1] that REMOS achieves very good recognition rates. However, the solution of the inner optimization problem causes a high computational load (i.e., real-time factors between 100 and 300). In this paper, we propose approximation techniques for the inner optimization problem which drastically reduce the computational complexity of REMOS. Since the decoding algorithm based on the proposed approximations ensures that the global optima of the inner optimization problem are found, it even increases the recognition accuracy.

This paper is structured as follows: The REMOS concept is concisely reviewed in Sec. 2. The proposed approximation is explained in Sec. 3, and connected digit recognition results are discussed in Sec. 4. Sec. 5 concludes the paper.

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under contract number KE 890/4-1.

 ISSN 2076-1465

2. REVIEW OF THE REMOS CONCEPT

The REMOS concept [6] extends conventional HMM-based speech recognition approaches by an RVM. Such an RVM can be interpreted as a statistical feature-domain description of all possible room impulse responses (RIRs) for arbitrary speaker and microphone positions in the target room. The RVM is represented in the logmelspec domain by an independent identically distributed (IID) matrix-valued Gaussian random process, where each column of the matrix corresponds to a certain delay m (in multiples of the frame shift) and each row of the matrix corresponds to a certain mel channel l. Therfore, the RVM is completely described by its mean matrix $\mu_{\boldsymbol{H}(0:M-1)} = \left[\mu_{\boldsymbol{H}(0)}, \dots \mu_{\boldsymbol{H}(M-1)}\right] \in \mathbb{R}^{L\times M}$ and its variance matrix $\sigma^2_{\boldsymbol{H}(0:M-1)} \in \mathbb{R}^{L\times M}$, where L denotes the number of mel channels and $0:M-1$ all frame delays from 0 to $M-1$.

In this paper, we stick to the following notational distinction: Every vector $\boldsymbol{v}$ without the explicit subscript "mel" is meant to be in the logmelspec domain whereas the corresponding vector $\boldsymbol{v}_{\text{mel}}$ denotes the melspec representation of $\boldsymbol{v}$, i.e.,

$$\boldsymbol{v} = \ln(\boldsymbol{v}_{\text{mel}}).$$

Furthermore, operators like "$=$", "$\leq$", "exp", "ln", and division applied to vectors are meant componentwise. Lower-case functions $v(k)$ are interpreted as realizations of the corresponding random process $V(k)$ denoted by upper-case letters.

In the REMOS framework we assume that the reverberant observation $\boldsymbol{x}_{\text{mel}}(k)$ is given in the melspec domain by

$$\boldsymbol{x}_{\text{mel}}(k) = \sum_{m=0}^{M-1} \boldsymbol{h}_{\text{mel}}(m,k) \odot \boldsymbol{s}_{\text{mel}}(k-m),$$

where $\boldsymbol{h}_{\text{mel}}(m,k)$ and $\boldsymbol{s}_{\text{mel}}(k-m)$ denote the output of the RVM and the output of the HMM network, respectively, and $\odot$ denotes the Hadamard product. For further simplification, we decompose

$$\boldsymbol{x}_{\text{mel}}(k) = \boldsymbol{h}_{\text{mel}}(0,k) \odot \boldsymbol{s}_{\text{mel}}(k) + \boldsymbol{a}_{\text{mel}}(k) \odot \boldsymbol{x}_{r,\text{mel}}(k), \quad (1)$$

where

$$\boldsymbol{x}_{r,\text{mel}}(k) = \sum_{m=1}^{M-1} \mu_{\boldsymbol{H}_{\text{mel}}(m)} \odot \boldsymbol{s}_{\text{mel}}(k-m) \quad (2)$$

is an approximation of the reverberant component

$$\sum_{m=1}^{M-1} \boldsymbol{h}_{\text{mel}}(m,k) \odot \boldsymbol{s}_{\text{mel}}(k-m),$$

and $\boldsymbol{a}_{\text{mel}}(k)$ is a vector of correction factors capturing the uncertainty of the corresponding approximation error. $\boldsymbol{a}(k)$ is a realization of the vector-valued Gaussian IID random process $\boldsymbol{A}(k)$, which is completely described by its mean vector $\mu_{\boldsymbol{A}} \in \mathbb{R}^L$ and its variance vector $\sigma^2_{\boldsymbol{A}} \in \mathbb{R}^L$.

By transforming (1) to the logmelspec domain, we obtain the following description for the observed reverberant feature vector sequence $\boldsymbol{x}(k)$:

$$\exp\big(\boldsymbol{x}(k)\big) = \exp\big(\boldsymbol{h}(0,k) + \boldsymbol{s}(k)\big) + \exp\big(\boldsymbol{a}(k) + \boldsymbol{x}_r(k)\big). \quad (3)$$

For recognition, an extended version of the Viterbi algorithm is employed to find the most likely path through the HMM network in connection with the RVM. At every step of the extended Viterbi algorithm, the Viterbi score is weighted by the outcome of the following inner optimization problem:

$$\max_{\boldsymbol{s}(k),\boldsymbol{h}(0,k),\boldsymbol{a}(k)} f_{\boldsymbol{S}(k)|Q(k)=j}\big(\boldsymbol{s}(k)\big) \cdot f_{\boldsymbol{H}(0)}\big(\boldsymbol{h}(0,k)\big) \cdot f_{\boldsymbol{A}(k)}\big(\boldsymbol{a}(k)\big)$$
$$\text{subject to (s.t.) } (3),$$

where $f_{\boldsymbol{H}(0)}$ and $f_{\boldsymbol{A}(k)}$ are the output densities of the RVM, and $f_{\boldsymbol{S}(k)|Q(k)=j}$ the output density of the j-th state of the HMM, each one modeled by a vector-valued Gaussian IID random process. The late reverberation $\boldsymbol{x}_r(k)$ is calculated by using estimates of $\boldsymbol{s}(k-m)$, $m = 1,...,M-1$, cf. (2), known from former Viterbi steps.

For solution, we decompose the optimization problem into the subproblems

$$\max_{\boldsymbol{x}_d(k),\boldsymbol{a}(k)} f_{\boldsymbol{X}_d(k)|Q(k)=j}\big(\boldsymbol{x}_d(k)\big) \cdot f_{\boldsymbol{A}(k)}\big(\boldsymbol{a}(k)\big) \quad (4)$$
$$\text{s.t. } \exp\big(\boldsymbol{x}(k)\big) = \exp\big(\boldsymbol{x}_d(k)\big) + \exp\big(\boldsymbol{a}(k) + \boldsymbol{x}_r(k)\big)$$

and

$$\max_{\boldsymbol{s}(k),\boldsymbol{h}(0,k)} f_{\boldsymbol{S}(k)|Q(k)=j}\big(\boldsymbol{s}(k)\big) \cdot f_{\boldsymbol{H}(0)}\big(\boldsymbol{h}(0,k)\big) \quad (5)$$
$$\text{s.t. } \boldsymbol{x}_d(k) = \boldsymbol{s}(k) + \boldsymbol{h}(0,k),$$

where $\boldsymbol{x}_d(k) = \boldsymbol{s}(k) + \boldsymbol{h}(0,k)$ denotes the direct sound component and $f_{\boldsymbol{X}_d(k)|Q(k)=j}$ is the corresponding probability density function.

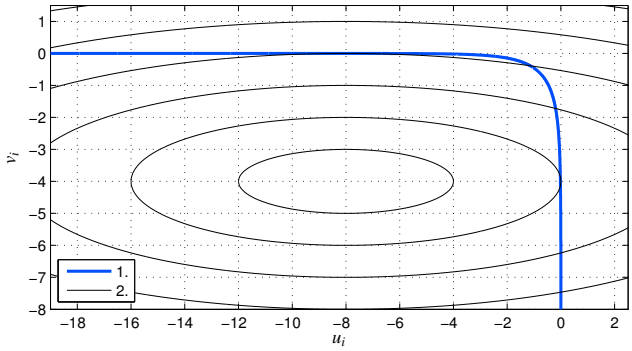
As (5) is explicitly solvable, we focus henceforth on the solution of (4). In [1], (4) has been solved by IPOPT (Interior Point OPTimizer), a general purpose software package for large-scale nonlinear optimization problems [7], causing the main computational load of the REMOS decoding algorithm.

3. EFFICIENT DECODING BY APPROXIMATION OF THE INNER OPTIMIZATION

3.1 Normalization

In this section, we are interested in the mathematical examination of (4) in order to derive approximations allowing an efficient solution. To get a deeper understanding of the structure of the problem, we drop the vector indices k and normalize some variables as follows:

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{x}_d - \boldsymbol{x} \\ \boldsymbol{v} &= \boldsymbol{a} + \boldsymbol{x}_r - \boldsymbol{x} \\ \mu_{\boldsymbol{U}} &= \mu_{\boldsymbol{X}_d} - \boldsymbol{x} \\ \mu_{\boldsymbol{V}} &= \mu_{\boldsymbol{A}} + \boldsymbol{x}_r - \boldsymbol{x} \end{aligned}$$



1.: exact constraint 2.: contour plot of J_i.

Figure 1: Illustration of the optimization problem (6) for one mel channel i.

Instead of maximizing $f_{X_d|Q=j}(\boldsymbol{x}_d) \cdot f_A(\boldsymbol{a})$, we minimize the quadratic functional

$$\begin{aligned} J(\boldsymbol{u},\boldsymbol{v}) &= -\ln(f_{X_d|Q=j}(\boldsymbol{u}+\boldsymbol{x}) \cdot f_A(\boldsymbol{v}-\boldsymbol{x}_r+\boldsymbol{x})) \\ &= \frac{1}{2}(\boldsymbol{u}-\mu_U)^T \operatorname{diag}(1/\sigma^2_{X_d})(\boldsymbol{u}-\mu_U) \\ &\quad + \frac{1}{2}(\boldsymbol{v}-\mu_V)^T \operatorname{diag}(1/\sigma^2_A)(\boldsymbol{v}-\mu_V) \\ &= \sum_{i=1}^{L} J_i(u_i,v_i) \end{aligned}$$

with

$$J_i(u_i,v_i) = \frac{(u_i-\mu_{U,i})^2}{2\sigma^2_{X_d,i}} + \frac{(v_i-\mu_{V,i})^2}{2\sigma^2_{A,i}}.$$

Finally, we obtain a normalized version of (4) which we consider for each mel channel $i = 1,...,L$ separately:

$$\begin{aligned} &\min_{u_i,v_i} J_i(u_i,v_i) \qquad (6) \\ &\text{s.t. } \exp(u_i)+\exp(v_i) = 1. \end{aligned}$$

An illustration of (6) is depicted in Fig. 1.

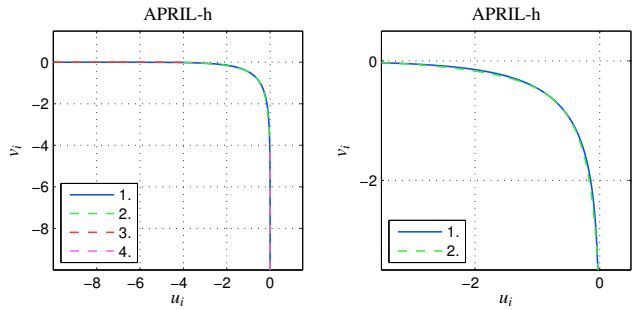
3.2 Hyperbolic Approximation

We now approximate the non-linear constraint (blue line in Fig. 1) in order to obtain an optimization problem which can be solved explicitly and efficiently. Therefore, we observe that the constraint in Fig. 1 tends very fast to its asymptotes $v_i = 0$ and $u_i = 0$ for $u_i \to -\infty$ and $v_i \to -\infty$, respectively, e.g., for $u_i < -4$, the difference between the constraint and its asymptote is less than 0.02. The segment for $u_i, v_i \geq -4$ can be very well approximated by a first-order hyperbola.
Hence, we approximate

$$\exp(u_i)+\exp(v_i) = 1 \qquad (7)$$

by

$$\begin{aligned} &\big(u_i \leq -4,\ v_i = 0\big) \ \vee\ \big(u_i = 0,\ v_i \leq -4\big) \\ &\quad \vee\ \big(-4 < u_i < 0,\ v_i = h(u_i)\big), \qquad (8) \end{aligned}$$



1.: $\exp(u_i)+\exp(v_i) = 1$, 3.: $u_i \leq -4,\ v_i = 0$,
2.: $-4 < u_i < 0,\ v_i = h(u_i)$, 4.: $u_i = 0,\ v_i \leq -4$.

Figure 2: Illustration of the exact constraint (7) and the approximated constraint (8). The different lines can hardly be distinguished in the diagram because of the high accuracy of the approximation.

where $h(u_i)$ is a first-order hyperbola of the form

$$h(u_i) = \frac{p}{u_i - q} + q$$

with

$$q = \frac{(\ln 2)^2}{4 - \ln 2}$$

and

$$p = 4q + q^2$$

chosen so that the three branches of (8) form a continous constraint, i.e., $h(0) = -4$ and $h(-4) = 0$, and (7) and (8) have the point $(-\ln 2, -\ln 2)$ in common. An illustration of (8) is depicted in Fig. 2. We finally obtain the following approximated optimization problem for each mel channel $i = 1,...,L$:

$$\begin{aligned} &O_i = \min_{u_i,v_i} J_i(u_i,v_i) \qquad (9) \\ &\text{s.t.: } (8), \end{aligned}$$

which we abbreviate *APRIL-h* (*h*yperbolic *APR*oximation of the *I*nner optimization problem in the *L*ogmelspec domain). In order to solve (9), we decompose it into three subproblems:

$$\begin{aligned} &A_i = \min_{u_i,v_i} J_i(u_i,v_i) \qquad (10) \\ &\text{s.t.: } u_i \leq -4,\ v_i = 0 \end{aligned}$$

as well as

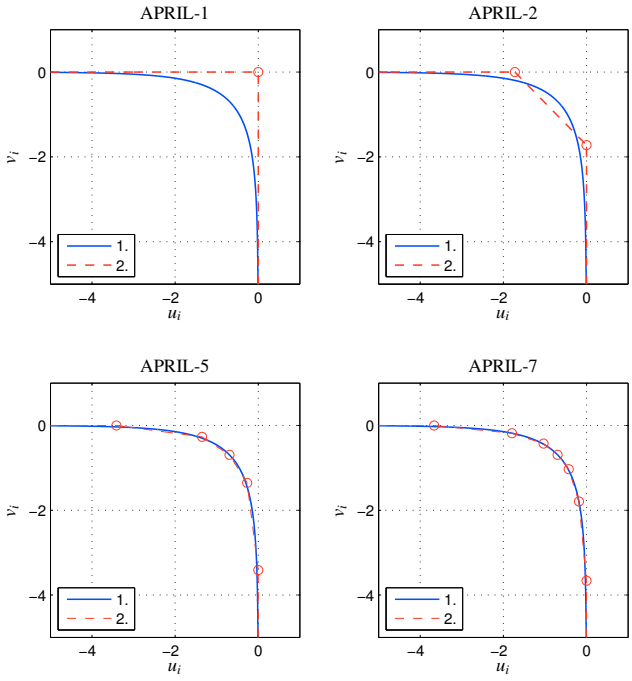
$$\begin{aligned} &B_i = \min_{u_i,v_i} J_i(u_i,v_i) \qquad (11) \\ &\text{s.t.: } -4 < u_i < 0,\ v_i = h(u_i) \end{aligned}$$

and

$$\begin{aligned} &C_i = \min_{u_i,v_i} J_i(u_i,v_i) \qquad (12) \\ &\text{s.t.: } u_i = 0,\ v_i \leq -4. \end{aligned}$$

Obviously, we have

$$O_i = \min\{A_i, B_i, C_i\}.$$



1.: exact constraint 2.: linear approximation.

Figure 3: Illustration of the exact constraint (7) and the piecewise linearly approximated constraint.

One can easily show that the solution of (10) fulfills

$$(u_i^*, v_i^*) = (0, \min\{\mu_{V,i}, -4\}).$$

Accordingly, the solution of (12) fulfills

$$(u_i^*, v_i^*) = (\min\{\mu_{U,i}, -4\}, 0).$$

The Lagrange system of (11)

$$\frac{\partial}{\partial u_i} J_i(u_i, h(u_i)) = 0$$

leads to a fourth-order polynomial equation which still can be explicitly solved [8].

3.3 Piecewise Linear Approximation

The introduced approximated optimization problem (9) can be solved much more efficiently than the original problem (6) with a negligible approximation error. However, solving a fourth-order polynomial equation still requires some computational load. To overcome this, we further reduce the complexity of (9) by approximating the exact constraint by a piecewise linear constraint. Fig. 3 shows several linear approximations for different number of nodes. The nodes were determined by a least-squares estimator. We abbreviate this type of approximation *APRIL-N*, where N is the number of nodes. For each segment of the piecewise linear approximation we obtain an optimization problem of the following type:

$$\min_{u_i, v_i} J_i(u_i, v_i) \qquad (13)$$
$$\text{s.t.: } r \leq u_i \leq s,\ v_i = m \cdot u_i + t,$$

where r, s, m and t are parameters determined by the least squares estimator. Since the problem has been normalized, these parameters are data-independent and therefore have to be determined only once. The Lagrange system of (13) leads to a linear equation which can be solved without any remarkable computational load.

3.4 APRIL versus IPOPT

To better exploit the advantages of IPOPT for large-scale problems it is not applied on the decomposed optimization problem (6) for each mel channel but on the overall problem (4) which has up to 2^L local optima. Therefore it is highly likely that IPOPT only finds a local optimum and not a global one. Since all APRIL approaches allow a closed form solution for the approximated problem it is guaranteed that they all find a global optimum.

4. EXPERIMENTS

Experiments with a connected-digit recognition task are carried out to analyze the performance of REMOS with the proposed optimization schemes.

4.1 Experimental Setup

The experimental setup is identical to [9]. Therefore, only the most important settings are summarized here. The REMOS-based recognizer is implemented by extending the decoding routines of HTK [10]. In the REMOS version of [1], the optimization problem (4) is solved by an interior-point line-search filter method implemented in IPOPT [7], an open-source software package for large-scale nonlinear optimization. Static log-melspec features with 24 mel channels calculated from speech data sampled at 20 kHz are used. 16-state word-level HMMs with single-Gaussian densities serve as clean-speech models. To get the reverberant test data, the clean-speech TI digits data are convolved with different RIRs measured at different loudspeaker and microphone positions in three rooms with the characteristics given in Table 1. A strict separation of training and test data is maintained in all experiments both for speech and RIRs. Each test utterance is convolved with an RIR selected randomly from a number of measured RIRs in order to simulate changes of the RIR during the test. The experiments are run on an Intel Q9550 with 2.83 GHz.

4.2 Experimental Results

Table 2 and table 3 compare the word accuracy and the real-time fators of REMOS to that of conventional

	Room A	Room B	Room C
Type	lab	studio	lecture room
T_{60}	300 ms	700 ms	900 ms
d	2.0 m	4.1 m	4.0 m

Table 1: Summary of room characteristics: T_{60} is the reverberation time, d is the distance between speaker and microphone.

	Room A	Room B	Room C
clean HMM	70.4	33.9	30.7
rev. HMM	89.1	82.3	69.9
REMOS	88.5	88.1	84.7
REMOS+APRIL-h	89.1	88.4	87.1
REMOS+APRIL-7	88.9	88.4	87.2
REMOS+APRIL-5	88.8	88.1	87.2
REMOS+APRIL-2	88.7	87.5	85.8
REMOS+APRIL-1	87.0	86.6	81.7

Table 2: Comparison of word accuracies in %.

recognizers with HMMs trained on clean speech and HMMs trained on reverberant speech. One can observe that the computational complexity of REMOS using the novel APRIL approaches is reduced by a factor of about 100 compared to the REMOS implementation of [1] using IPOPT which solves the exact optimization problem (4). This can be explained by the fact that APRIL directly calculates the closed-form solution of the approximated problem whereas IPOPT applies a generic iterative method to the exact problem which entails a certain convergence time. The recognition rates of REMOS+APRIL-h, APRIL-5 and APRIL-7 are even higher than those of the original REMOS version. This confirms that finding a global optimum for the inner optimization problem leads to improved recognition rates. The improved recognition rates for the global optimum also confirm the validity of the reverberation model and of the formulation of the optimization problem. Compared to the computional load of the memory access for calculating $\boldsymbol{x}_r$, cf. (2), the complexity of APRIL-1, APRIL-2 and APRIL-5 is negligible.

5. SUMMARY AND CONCLUSIONS

The REMOS concept for robust distant-talking speech recognition according to [1] has been modified to drastically reduce computational demands, so that real-time implementations are within reach. Approximation techniques for the inner optimization problem, which has to be solved in each iteration of the Viterbi algorithm, allow an efficient solution and implementation. Connected digit recognition experiments confirm that the computational complexity is significantly reduced with a remarkable increase of the recognition accuracy. Future work will include extending the concept to the MFCC domain, to Gaussian mixture densities, and dynamic features as well as optimizing the storage policy for further reducing the RTF.

	Room A	Room B	Room C
clean HMM	< 0.1	< 0.1	< 0.1
rev. HMM	< 0.1	< 0.1	< 0.1
REMOS	265	178	196
REMOS+APRIL-h	1.7	2.5	2.9
REMOS+APRIL-7	1.0	1.7	2.4
REMOS+APRIL-5	0.9	1.6	2.1
REMOS+APRIL-2	0.9	1.6	2.1
REMOS+APRIL-1	0.9	1.6	2.1

Table 3: Comparison of real-time factors (RTF).

REFERENCES

[1] A. Sehr, R. Maas, and W. Kellermann, "Model-based dereverberation in the logmelspec domain for robust distant-talking speech recognition," accepted for *ICASSP 2010*, Dallas, Texas, USA, March 14-19. 2010.

[2] *Special Issue on Processing Reverberant Speech: Methodologies and Applications*, for the IEEE Transactions on Audio, Speech and Language Processing, to appear.

[3] C. K. Raut, T. Nishimoto, and S. Sagayama, "Model adaptation by state splitting of HMM for long reverberation," in *Proc. ICSLP 2005*, Lisbon, Portugal, September 4-8. 2005, pp. 277–280.

[4] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," in *Proc. ICASSP 1999*, Phoenix, Arizona, USA, March 15-19. 1999, pp. 449–452.

[5] T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, "Using artificially reverberated training data in distant-talking ASR," in *Proc. TSD 2005*, Karlovy Vary, Czech Republic, September 12-15. 2005, pp. 226–229.

[6] A. Sehr and W. Kellermann, *Towards robust distant-talking automatic speech recognition in reverberant environments*, in E. Hänsler and G. Schmidt, editors, Topics in Speech and Audio Processing in Adverse Environments, pp. 679–728, Berlin: Springer, 2008.

[7] A. Wächter and L.T. Biegler, "On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, no. 1, pp. 25–57, 2006.

[8] L. Euler, *Elements of Algebra*, Tarquin Publications, 2006.

[9] A. Sehr, M. Zeller, and W. Kellermann, "Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain," in *Proc. INTERSPEECH 2006*, Pittsburgh, PA, USA, September 17-21. 2006, pp. 769–772.

[10] "HTK webpage," http://htk.eng.cam.ac.uk/.