Blind Model Selection for Automatic Speech Recognition in Reverberant Environments

Laurent Couvreur* and Christophe Couvreur

March 22, 2004

Abstract

This communication presents a new method for automatic speech recognition in reverberant environments. Our approach consists in the selection of the best acoustic model out of a library of models trained on artificially reverberated speech databases corresponding to various reverberant conditions. Given a speech utterance recorded within a reverberant room, a Maximum Likelihood estimate of the fullband room reverberation time is computed using a statistical model for short-term log-energy sequences of anechoic speech. The estimated reverberation time is then used to select the best acoustic model, *i.e.*, the model trained on a speech database most closely matching the estimated reverberation time, which serves to recognize the reverberated speech utterance. The proposed model selection approach is shown to improve significantly recognition accuracy for a connected digit task in both simulated and real reverberant environments, outperforming standard channel normalization techniques.

Keywords: room reverberation, maximum likelihood estimation, automatic speech recognition

Author Addresses:

Laurent Couvreur	Christophe Couvreur
Multitel - TCTS	Speech & Language Technology Division
Faculté Polytechnique de Mons	Scansoft, Inc.
1 Avenue Copernic	32 Guldensporenpark
B-7000 Mons	B-9820 Merelbeke
Belgium	Belgium
Tel: +32 65 374775	Tel: +32 9 239 81 64
Fax: +32 65 374729	Fax: +32 9 239 80 01
Email: couvreur@multitel.be	Email: Christophe.Couvreur@scansoft.com

*This work was supported in part by F.N.R.S. (Fonds National de la Recherche Scientifique) with a F.R.I.A grant (Fonds pour la formation à la Recherche dans l'Industrie et l'Agriculture, Belgium).

1 Introduction

During the past decade, automatic speech recognition (ASR) has been successfully deployed in desktop applications (*e.g.*, dictation systems) and in interactive telephone-based systems (*e.g.*, voice portals). These systems typically operate with a close-talking microphone in quiet environments. Nowadays, we observe a growing interest for voice-controlled devices which can be used in more natural conditions, the so-called *hands-free* devices [1]. However, there still exist important obstacles to the mass development of such devices. The poor accuracy of the current ASR systems in realistic hands-free operating conditions is the main one.

Today's state-of-art ASR systems are based on stochastic models of speech acoustics. That is, they extract a set of significant acoustic features from the speech signal and use statistical models to represent the distribution of these features for speech items such as words, syllables or phonemes. Recognition of an unknown utterance is then treated as a statistical pattern recognition problem. These statistical models, often referred to as acoustic models, have to be trained on large speech databases. Unfortunately, the performance of the ASR systems degrade dramatically when it is presented with speech having different acoustic features than the ones statistically modeled during the training procedure. For hands-free devices, the speech signal propagates from the speaker to a distant microphone through an unknown acoustic environment. In realistic acoustic environments, the speech signal is contaminated by competing noise sources and distorted by room reverberation before reaching the microphone. Since the ASR acoustic models are usually trained on noise-free anechoic speech, they are unlikely to match the distribution of acoustic features of the recorded speech in a noisy reverberant environment.

In this communication, we are primarily concerned by the effect of room reverberation. When an acoustic signal (a speech signal) is generated in a reverberant room, it follows multiple paths from the source (the speaker) to the receiver (the microphone). While a version of the source signal propagates directly to the receiver through the air, other delayed versions arrive after bouncing back and forth on the reflective surfaces of the room. The phenomenon severely affects the spectral characteristics of the acoustic signal picked up at the microphone. It globally results in a "smearing" effect of the acoustic energy forward in time. Figure 1 shows the waveforms and the corresponding spectrograms of an anechoic speech utterance recorded with a close-talking microphone and its reverberant version recorded with a distant-talking microphone in a typical reverberant room.

The distortion caused by room reverberation is especially harmful for ASR systems [2, 3, 4]. Several approaches have been proposed to reduce the discrepancy between the training conditions (close-talking/anechoic speech) and the operating conditions (distant-talking/reverberated speech) due to room reverberation. These methods can be roughly separated into three categories:

Speech Enhancement A first approach consists in attempting to recover an "enhanced" speech signal close to the anechoic one by processing its reverberated version while keeping the acoustic models unchanged. These methods may be further categorized into single-microphone

methods [5, 6, 7, 8, 9] and multi-microphone methods [10, 11, 12, 13, 14]. While these methods can produce high quality enhanced speech, they are generally computationally demanding and may require solving ill-conditioned mathematical problems [15, 16].

Model Adaptation Other methods propose to reduce the mismatch between the training and the operating conditions by adapting the acoustic models from observed reverberated speech. Many efficient adaptation schemes [17, 18, 19, 20, 21, 22] have been proposed to address the problem of ASR in reverberant environments. However, they generally require a significant amount of adaptation data, thereof limiting their practical use in realistic conditions where the acoustic environment is rapidly changing.

Robust Features The last class of methods includes all the preprocessing parts of the ASR systems, the so-called front-ends, that extract acoustic features of speech insensitive to channel distortion [23, 24, 25, 26, 27, 28, 29]. These methods have the advantage of being easy to implement. They are cost-effective since there is no need to adapt the acoustic models given reverberated observations or to restore the speech signal prior to computing the acoustic features. However, they often perform poorly in moderately to highly reverberant environments.

In this communication, we investigate a new technique that can be somewhat regarded as a model adaptation method. In our approach, a library of acoustic models are trained separately on speech databases corresponding to various reverberant conditions. During operation, the best model is selected and used to recognize the incoming reverberated speech. The best model is actually the model that most closely matches the operating reverberant conditions. Unlike in classical adaptation techniques, the speech data to be recognized are not used to adapt the acoustic model but simply to select one from the library. This method is fast since it requires little data for "adapting" the acoustic model by selection from the library. It is also computationally efficient since the model "adaptation" is actually done beforehand during the training procedure. In order to implement this approach, two issues have to be addressed: the procedure for creating the library of acoustic models and the procedure for selecting a model from the library during operation,

Training Procedure Ideally, a training database should be collected for every possible operating reverberant environment where the ASR system is to be deployed. Hence, numerous large speech databases would need to be collected in order to cover most of reverberant conditions. The problem is further compounded if the ASR system must be speaker-independent and if several languages must be supported. The database collection effort rapidly becomes unmanageable.

Alternatively, the reverberated speech databases may be obtained by adequately reverberating an anechoic speech database. It is commonly assumed that the transmission of an acoustic signal from a source to a microphone within a reverberant enclosure may be modeled by a

linear filter. Under this assumption, a reverberated speech database may be obtained by convolving an anechoic speech database with a room impulse response measured between the source and the microphone in the operating reverberant environment [30, 17, 31]. However, this approach ignores that room impulse responses are highly dependent on the acoustic characteristics of the reverberant enclosure (*e.g.*, a room impulse response changes when the temperature and the humidity vary or when air currents shift [32]), on the geometric characteristics of the room (*e.g.*, a room impulse response changes when doors are opened or closed), on the source and the microphone locations (*e.g.*, a room impulse response changes when the speaker moves around [15]), etc. It is thus extremely hard to guarantee that the measured room impulse response would exactly match the room impulse response of the operating reverberant environment. There is a risk that the ASR system will then "over-fit" the reverberant conditions observed while measuring the room impulse response, leading to disappointing results when the operating conditions change, even slightly [3].

To resolve this problem, we proposed in [3] to use a "randomized" synthetic reverberating filter instead of a measured room impulse response to obtain reverberated speech databases. The impulse response for the synthetic reverberating filter is generated multiple times randomly during the reverberation process of the anechoic speech database. It is constrained to match a high-level, perceptually meaningful, acoustic property of the operating reverberant environment, namely the fullband reverberation time T_{60} [33]. The randomization of the low-level details of the impulse response is expected to model the effects of possible speaker/microphone movements and of small variations in enclosure geometry or acoustic characteristics, consequently avoiding over-fitting the acoustic models to a particular room impulse response. Artificially reverberated speech databases (and the acoustic models thereof) can be generated for various reverberation times to build the desired library of acoustic models. This approach requires neither collecting large speech databases nor measuring room impulse responses. While the assumption that the fullband reverberation time T_{60} is sufficient to capture the properties of a reverberant environment may seem overly simplistic, it has been found to yield good results in the framework of ASR. The complete training procedure is presented in details in [3].

Selection Procedure Given a reverberated speech utterance to be recognized, we want to select the best acoustic model out of the library generated with the training procedure described above. In our approach, it is sufficient to estimate the fullband reverberation time T_{60} for the speech utterance and then select an acoustic model accordingly. This communication describes an algorithm for the blind estimation of T_{60} given a speech utterance recorded with a distant-talking microphone in a reverberant environment. The proposed algorithm is inspired by the Stochastic Matching concept introduced by Sankar and Lee in [34]. Its basic principles can be briefly summarized as follows.

The impact of room reverberation is defined in the short-time log-energy domain by a parametric model whose parameters are related to the reverberation time. Although this model relies on a strong assumption, namely that room reverberation can be entirely characterized by the fullband reverberation time T_{60}, it allows us to obtain a mathematically tractable model and, as will be seen later, it works satisfyingly in practice. Given a reverberated speech utterance and its corresponding short-term log-energy sequence, the estimation algorithm computes a Maximum Likelihood (ML) estimate of the parameters of the distortion model, and an estimate of the reverberation time is derived. This algorithm requires a statistical model of short-term log-energy sequences for anechoic speech that must be trained beforehand on an anechoic speech database.

In the next section, we define the simplified room reverberation model used for the selection procedure. Then, the algorithm for the blind estimation of T_{60} is presented in section 3. In section 4, we report experimental results for the recognition of connected digit sequences. For these experiments, our model selection procedure is tested with a library of acoustic models built with artificially reverberated speech databases. Results are given for both simulated and real reverberant environments. In the former case, test material is obtained by simulating a rectangular room with the Image Method [35, 36]. In the latter case, test recordings have been collected in various real reverberant enclosures. The model selection approach is compared with standard channel compensation techniques [23, 24]. Although the hands-free ASR problem is addressed only partially in this communication (*i.e.*, reverberant but noise-free conditions are considered), the method shows promising results. The results obtained in section 4 allow us to identify limitations of the current approach. Conclusions and directions for future research to address these shortcomings are given in section 5.

2 Room Reverberation Model

The propagation between a source and a microphone within a reverberant enclosure is classically modeled in the time domain by a linear filter. A reverberated speech waveform y_n is obtained by convolving its anechoic version x_n with a causal room impulse response h_n,

$$y_n = x_n * h_n.$$

As mentioned above, room impulse responses are highly dependent on the acoustic and geometric characteristics of the room, and on the source and microphone locations. However, under the diffuse sound field assumption [33], it can be shown that the ensemble average ε_n of the squared room impulse responses is a decaying exponential:

$$\varepsilon_n \stackrel{\triangle}{=} < h_n^2 >= \varepsilon_0 e^{-kn}, \tag{1}$$

where the damping constant k is related to the fullband reverberation time T_{60} [s] of the reverberant room by

$$k = \log 10^6 / (T_{60} \times F_s) \tag{2}$$

with F_s [Hz] standing for the sampling frequency. Equation (2) is based on the classical definition of the reverberation time [33]: T_{60} is the time interval in which the sound energy within a reverberant room reaches one millionth (*i.e.*, -60 dB) of its initial value once the sound source has been interrupted. These equations suggest that it is possible to generate a synthetic room impulse response matching a given reverberation time by modulating a zero-mean random sequence with a decaying exponential [37]. This forms the basis for the technique proposed in [3] to generate artificially reverberated speech databases by filtering an anechoic speech database in order to train acoustic models in various reveberant conditions, *i.e.*, for various reverberation times.

Room reverberation is fully described by the room impulse response between the speaker and the microphone. In order to estimate the reverberation time T_{60} for selecting an acoustic model, one can propose to identify blindly the room impulse response from recorded reverberated speech, and then to compute the reverberation time from the estimated impulse response, for instance, using Schroeder's method [33]. But since the blind identification of a room impulse response is a difficult ill-conditioned problem, we propose to use instead a coarser model of room reverberation to simplify the T_{60} estimation problem. We decide to model the impact of room reverberation on short-term energy sequences instead of on time-domain speech signals. The short-term energy sequence W_m of an anechoic speech signal x_n (the speaker voice) is obtained by slicing it into (possibly overlapping) frames, and by computing the energy of every frame. For the m-th frame, we have

$$W_m \triangleq \frac{1}{N_w} \sum_{n=mN_r}^{mN_r+N_w-1} x_n^2, \tag{3}$$

where $N_w \triangleq T_w \times F_s$ and $N_r \triangleq F_s/F_r$ with T_w [s] and F_r [Hz] denoting the frame length and the frame rate, respectively. Likewise, we define Z_m as the short-term energy sequence of the reverberated speech signal y_n (the microphone signal). Under the assumption that the sound field is diffuse and that equation (1) is valid, the energy of the m-th frame W_m will smear forward in time, its contribution to the following frames decaying exponentially. We assume that this smearing effect in the short-term energy domain can be modeled by a first order auto-regressive (AR) filter:

$$\alpha_0 Z_m = W_m - \alpha_1 Z_{m-1}, \tag{4}$$

with $\alpha_0 > 0$ and $\alpha_1 < 0$.

Figure 2 summarizes the reverberation model for short-term log-energy sequences and its relation to the original signals. Figure 3 gives an example of an anechoic speech utterance x_n and its reverberated version y_n recorded with a close-talking microphone and a distant-talking micro-

phone, respectively, in a typical reverberant room. The figure also shows the distortion on the corresponding short-term log-energy sequences $X_m \triangleq 10\log_{10} W_m$ and $Y_m \triangleq 10\log_{10} Z_m$ computed after proper normalization of the speech signals. The short-term energy sequences are represented in the logarithmic domain instead of the linear domain because of better conditioned dynamics there. In figure 3, we observe that the decay of Y_m from peak to valley is almost linear. The decay is thus exponential in the short-term linear-energy domain, which validates the model postulated in equation (4). Like in equation (2), the decaying rate of the short-term energy can be related to the reverberation time T_{60} by

$$T_{60} = \log 10^6 / (-\log(-\alpha_1) \times F_r). \tag{5}$$

3 Blind T_{60} Estimation Algorithm

We present here an algorithm for blindly estimating the parameters (α_0, α_1) of the reverberation model given a reverberated utterance y_n. Once α_1 has been estimated, T_{60} of the reverberant room can be derived via (5). The algorithm is blind because no other information (*e.g.*, silence/speech segmentation or close-talking reference of the speech signal) other than a speech utterance recorded with a distant-talking microphone are needed. First, a statistical model for short-term log-energy sequences of anechoic speech utterances is introduced. Next, an estimation algorithm using this model is described. Finally, some results to demonstrate its convergence properties are reported.

3.1 Model for Speech Production

Typical short-term log-energy sequences for anechoic speech are non-stationary and characterized by two states, called the silence and speech states. Furthermore, successive values in a given state are undoubtedly not statistically independent (see figure 3.(c)). Hence, we decide to model short-term log-energy sequences for anechoic speech by a 2-state one-dimensional Linear Predictive Hidden Markov Model (LP-HMM) [38, 39]. In this model, a short-term log-energy sequence X_m for anechoic speech is generated by processing an emission sequence E_m with an AR filter of order P,

$$\beta_0(s_m) X_m = E_m - \sum_{p=1}^{P} \beta_p(s_m) X_{m-p} \tag{6}$$

whose coefficients $\beta_p(s_m)$, $p = 0, \ldots, P$, vary in time according to the evolution of a hidden Markov chain. The emissions E_m are assumed to be conditionally independent given the LP-HMM state sequence s_m and to have a Gaussian distribution with mean μ_i and standard deviation σ_i for state $s_m = i$, $i = \{0, 1\}$. To complete our model, we define the transition probabilities $a_{ij} \triangleq P[s_m = j | s_{m-1} = i]$.

Figure 4 illustrates a 2-state one-dimensional LP-HMM with AR filters limited to first order (*i.e.*,

$P = 1$) which is used in this work for modeling short-term log-energy sequences of anechoic speech utterances. Note that the LP-HMM is defined for log-energy sequences instead of linear-energy sequences because these data have better conditioned dynamics, and their statistical distributions are easier to model. All the parameters of the LP-HMM can be estimated by an Expectation-Maximization (EM) algorithm [40, 39]. Table 1 gives the parameters of the LP-HMM used in this work. The model has been trained on a noise-free part of the AURORA speech database [41]. Note that each speech utterance is normalized with respect to its energy prior to computing its short-term log-energy sequence.

3.2 Algorithm Description

The proposed algorithm for estimating T_{60} is inspired by the Maximum Likelihood (ML) Stochastic Matching paradigm introduced by Sankar and Lee [34]. The Stochastic Matching method assumes that some features X are observed after a parametric distortion $F(\cdot \mid \boldsymbol{\alpha})$ such that only the distorted features $Y = F(X \mid \boldsymbol{\alpha})$ are available. It provides a framework for deriving a ML estimate of the distortion parameters $\boldsymbol{\alpha}$ given some observed features Y and a statistical model for the unobserved features X.

Here, we consider the short-term log-energy sequences of reverberated speech utterances (microphone signal) as the observed features. The short-term log-energy sequences of the corresponding anechoic speech utterances (speaker voice) are the unobserved features. The distortion model is the room reverberation model in the short-term log-energy domain of figure 2. Hence, the distortion parameters are the coefficients (α_0, α_1) of the first-order AR filter of equation (4). The statistical model of the unobserved features is the 2-state LP-HMM described in section 3.1.

Let $Y_{0,M} = \{Y_0, \ldots, Y_M\}$ denote the finite-length short-term log-energy sequence of a reverberated speech utterance and define $s_{0,M} = \{s_0, \ldots, s_M\}$ as the corresponding hidden state sequence of the underlying LP-HMM. We want to compute an estimate of the distortion parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ based on the Stochastic Matching paradigm. That is, we want to solve the ML estimation problem

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \log p(Y_{0,M} \mid \boldsymbol{\alpha}) \tag{7}$$

where $p(Y_{0,M} \mid \boldsymbol{\alpha})$ denotes the likelihood of the data $Y_{0,M}$ given the distortion parameters $\boldsymbol{\alpha}$. Since the underlying state sequence is unknown, we have to solve a ML problem for incomplete data. We naturally resort to an Expectation-Maximization (EM) algorithm [40] to derive a solution. The EM algorithm consists in a two-step iterative procedure.

E-step – Given current estimates $\boldsymbol{\alpha}^{(\ell)}$ of the distortion parameters at iteration ℓ and some observations $Y_{0,M}$, compute the conditional expectation, also known as the auxiliary function:

$$Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{(\ell)}) \stackrel{\triangle}{=} \mathrm{E}\left[\log p(Y_{0,M}, s_{0,M} \mid \boldsymbol{\alpha}) \mid Y_{0,M}, \boldsymbol{\alpha}^{(\ell)}\right]. \tag{8}$$

M-step – Find new estimates $\boldsymbol{\alpha}^{(\ell+1)}$ of the distortion parameters by maximizing equation (8) with respect to $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{(\ell+1)} = \arg\max_{\boldsymbol{\alpha}} Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^{(\ell)}). \tag{9}$$

Although numerical optimization techniques can be used to solve (9), we prefer deriving an explicit solution, the so-called re-estimation formulae. We first compute the first derivative of the auxiliary function with respect to $\boldsymbol{\alpha}$. Next, we solve for its zeros to obtain the new estimates $\boldsymbol{\alpha}^{(\ell+1)}$. They are then substituted in (8) to iterate the procedure, unless convergence is reached.

In order to derive close-form re-estimation formulae, we first apply the Bayes rule and note that the state sequence does not depend on the distortion parameters. Hence, we can equivalently solve

$$\boldsymbol{\alpha}^{(\ell+1)} = \arg\max_{\boldsymbol{\alpha}} \mathrm{E}\left[\log p(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha}) \mid Y_{0,M}, \boldsymbol{\alpha}^{(\ell)}\right]. \tag{10}$$

In appendix A, we derive an approximation of the transformation $E_{0,M} = G(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha})$ between a finite length sequence of observations $Y_{0,M}$ and the corresponding sequence of emissions $E_{0,M}$ given a state sequence $s_{0,M}$. The approximation allows us to express the likelihood function for the observations with respect to the likelihood function for the emissions [42] as

$$p(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha}) = p(E_{0,M} \mid s_{0,M}, \boldsymbol{\alpha})\, |\mathbf{J}|, \tag{11}$$

where $|\mathbf{J}|$ denotes the determinant of the Jacobian matrix $\mathbf{J}$ of the transformation $G(\cdot)$. In appendix A, we also show that $|\mathbf{J}|$ depends only on α_0 (see equation (32)). By proper normalization of the energy levels of the speech utterances during the training procedure of the LP-HMM and during the estimation of T_{60}, we can constrain $\alpha_0 = 1$. The maximization problem is then reduced to

$$\alpha_1^{(\ell+1)} = \arg\max_{\alpha_1} \mathrm{E}\left[\log p(E_{0,M} \mid s_{0,M}, \alpha_1) \mid Y_{0,M}, \alpha_1^{(\ell)}\right]. \tag{12}$$

By definition of the LP-HMM in section 3.1, the emissions $E_{0,M}$ are conditionally independent, and equation (12) becomes,

$$\alpha_1^{(\ell+1)} = \arg\max_{\alpha_1} \sum_{m=0}^{M} \mathrm{E}\left[\log p(E_m \mid s_m, \alpha_1) \mid Y_{0,M}, \alpha_1^{(\ell)}\right]. \tag{13}$$

The expectation operator can be factored out at ℓ-th iteration by introducing the *a posteriori* state probabilities $\gamma_{m,i}^{(\ell)} \triangleq P[s_m = i \mid Y_{0,M}, \alpha_1^{(\ell)}]$, $i = \{0, 1\}$, yielding

$$\alpha_1^{(\ell+1)} = \arg\max_{\alpha_1} \sum_{m=0}^{M} \sum_{i=0}^{1} \gamma_{m,i}^{(\ell)}\, \log p(E_m \mid s_m = i, \alpha_1). \tag{14}$$

Given the parameters of the LP-HMM, the *a posteriori* state probabilities can be efficiently com-

puted for any sequence of observations with the Forward-Backward algorithm [39]. Using the assumption that the emissions have Gaussian distributions and that the distribution parameters are state-dependent, we find

$$\alpha_1^{(\ell+1)} = \arg\max_{\alpha_1} \sum_{m=0}^{M} \sum_{i=0}^{1} \gamma_{m,i}^{(\ell)} (E_m - \mu_i)^2 / \sigma_i^2, \tag{15}$$

where constant terms have been left out for ease of notation. Next, we can use the linear approximation (24) derived in appendix A and make $\alpha_1^{(\ell+1)}$ appear explicitly:

$$\alpha_1^{(\ell+1)} = \arg\max_{\alpha_1} \sum_{m=0}^{M} \sum_{i=0}^{1} \gamma_{m,i}^{(\ell)} \left(\sum_{p=0}^{1} \beta_p(s_m) \left[\sum_{k=0}^{1} \frac{\alpha_k^{(\ell+1)}}{\xi W_m^{(\ell)}} Z_{m-p-k} + X_m^{(\ell)} - 1/\xi \right] - \mu_i \right)^2 / \sigma_i^2 \tag{16}$$

with $\xi = \log(10)/10$. The re-estimation formula for α_1 is finally obtained by setting the first-order derivative of the right-hand term of (16) with respect to $\alpha_1^{(\ell+1)}$ to zero. Note that this first-order derivative is linear for α_1. We finally get

$$\alpha_1^{(\ell+1)} = (-p^{(\ell)} - q^{(\ell)}) / r^{(\ell)}, \tag{17}$$

with $p^{(\ell)}$, $q^{(\ell)}$ and $r^{(\ell)}$ defined by

$$c_{m,i}^{(\ell)} \triangleq \sum_{p=0}^{1} \beta_p(i) Z_{m-p-1} / \xi W_{m-p}^{(\ell)}, \tag{18}$$

$$r^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(c_{m,i}^{(\ell)} \right)^2, \tag{19}$$

$$q^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(\sum_{p=0}^{1} \frac{\beta_p(i) Z_{m-p}}{\xi W_{m-p}^{(\ell)}} \right) c_{m,i}^{(\ell)}, \tag{20}$$

$$p^{(\ell)} \triangleq \sum_{m=0}^{M} \sum_{i=0}^{1} \frac{\gamma_{m,i}^{(\ell)}}{\sigma_i^2} \left(\sum_{p=0}^{1} \beta_p(i) (X_{m-p}^{(\ell)} - \frac{1}{\xi}) \; - \; \mu_i \right) c_{m,i}^{(\ell)}. \tag{21}$$

The resulting iterative algorithm for the ML estimation of the distortion parameter α_1, and therefrom the reverberation time T_{60} via equation (5), is summarized in table 2. Note that the Viterbi approximation [38, 39] may be postulated, taking into account only the most likely state path $s_{0,M}$ instead of considering all the possible state sequences via the Forward-Backward algorithm. In this case, the *a posteriori* state probabilities are constrained to be null except for one state at every time instant. This results in a computationally more efficient but suboptimal estimation algorithm.

3.3 Convergence Results

The proposed algorithm for the estimation of T_{60} has been derived in the EM framework. We can thus expect it to be convergent, if not for the approximations that had to be made. In this section, we do not mathematically prove the convergence of the estimation algorithm but rather report some results which demonstrate its convergence properties in practice.

First, we show by an example the convergence of the estimation algorithm for various values of the distortion parameter α_1. For every estimation trial, 256 observations are produced by applying the distortion model (see section 2) to emissions drawn from the LP-HMM for anechoic speech (see section 3.1). Actually, α_0 is set equal to 1 and five values of α_1 ranging from -0.45 to -0.85 are tested. The estimation algorithm is initialized with $\alpha_1^{(0)} = -0.95$. The iterative estimation procedure is stopped when two successive estimates do not differ by more than $\delta = 10^{-4}$ or if a maximum number $\ell_{\max} = 128$ of iterations is reached. Figure 5 shows the evolution of the estimate $\alpha_1^{(\ell)}$ as a function of the number ℓ of iterations. The estimation algorithm clearly converges and an accurate estimate of the distortion parameter α_1 is obtained after a few iterations only.

Next, we run a series of Monte-Carlo experiments. Observations are again generated by applying the distortion model (α_1 varying from -0.45 to -0.85) to emissions of the anechoic speech LP-HMM ($M + 1$ varying from 8 to 512). The initialization and the stopping criterion are as described above. For every setup, 1024 Monte-Carlo trials are performed and the mean squared error (MSE) between the estimated distortion parameter $\alpha_1^{(\ell+1)}$ and the true one α_1 is computed. Figure 6 shows that the estimation algorithm is consistent, that is, the MSE value decreases when the number of observations increases. We clearly observe that the more observations the better the estimate, especially for high values of the distortion parameter α_1.

4 Speech Recognition Results

In this section, we report some results of our model selection approach for automatic speech recognition in reverberant environments. We first describe the experimental setup. Next, we give ASR results for both simulated reverberant speech and for real data recorded in reverberant enclosures.

4.1 Experimental Setup

The speech database used in this work comes from the noise-free part of the AURORA [41] corpus. It consists of connected digit sequences and it is divided into a training set of 8840 utterances and a test set of 1001 utterances, pronounced by 110 speakers and 104 other speakers, respectively. Recognition experiments are performed with a phoneme-based hybrid system relying on the Hidden Markov Model / Multilayer Perceptron (HMM/MLP) paradigm [43]. That is, a MLP is used for the acoustic modeling stage. The MLP is fed with acoustic features computed from 30ms long/10ms overlapping frames of speech signal sampled at 8kHz. For every frame, it estimates the phoneme

a posteriori probabilities. In the ASR experiments, three types of acoustic features are used: Mel-warped frequency cepstral coefficients (MFCC), MFCC with cepstral mean subtraction (CMS) [23] and logRASTA-PLP [24] coefficients. The last two front-ends are known to be robust to channel distortion. Speech decoding is done by Viterbi search, with neither pruning nor grammar constraints.

As described previously, we need to build a library of acoustic models (MLP's) trained on speech databases corresponding to various reverberant conditions. We first train a MLP on anechoic speech. Then, we train eight MLP's on artificially reverberated training sets. Training material is obtained with the reverberation process described in [3] for eight values of T_{60} varying uniformly from 200ms to 1600ms. All MLP's are obtained with the classical supervised back-propagation algorithm [43].

Once the library of acoustic models have been generated, we use them to recognize various reverberant test sets by selecting the most appropriate model. The selection is based on the estimation of the fullband reverberation time T_{60} by the method described in section 3. Like all iterative methods, our T_{60} estimation algorithm requires an initial estimate $\alpha_1^{(0)}$ (see section 3.2), and a stopping criterion. In practice, the reverberation time is lower than 2 seconds in most real reverberant enclosures. This upper bound is used to compute the initial estimate via equation (5), that is, $\alpha_1^{(0)} = -0.933$. The stopping criterion is the one described in section 3.3. The LP-HMM modeling short-term log-energy sequences (see section 3.1) is trained on a 64-utterance subset of the anechoic training set.

4.2 Results on Simulated Speech Data

Our model selection method is first tested on simulated reverberant speech. The test sets are obtained by convolving anechoic speech with room impulse responses computed by the Image Method [35, 36]. The main advantage of the room simulation approach is the flexibility and control it offers on the reverberation time. Various test sets are generated within a 9m long $\times$ 6m wide $\times$ 4m high simulated room. The microphone is located 2m away from the speech source. The wall absorption coefficients are set such that T_{60} is ranging from 200ms to 1600ms.

First, we report cross-testing results (see table 3). We see that the lowest word error rate (WER), *i.e.*, the sum of the substitution (SUB), deletion (DEL) and insertion (INS) error rates, is always achieved by the acoustic model most closely matching the testing conditions (along main diagonal). Note that, as could have been expected, WER increases for the matching acoustic model when the reverberation becomes stronger.

Next, we apply our model selection approach by blind estimation of T_{60}. For all test sets, each reverberated speech utterance is processed prior to its recognition: the speech signal amplitude is normalized, the short-term log-energy sequence is computed for $T_w = 30$ms and $F_r = 100$Hz, T_{60} is estimated and the most closely matching MLP of the library is activated. Note that the blind T_{60}

estimation algorithm becomes ill-conditioned for anechoic speech. But since real rooms are never anechoic, this is not a problem in practice.

Table 4 reports the performance of the selection method in terms of median absolute value errors (MAVE) for T_{60} estimation and confusion rates (CR), *i.e.*, wrong acoustic model selection rates. The estimation algorithm performs satisfactorily and the most closely matching MLP is very often selected. The proposed model selection method approaches the recognition performance of the "Oracle" method (see table 5) for which T_{60} is assumed to be known in advance and the best model is always selected. Clearly, the model selection method outperforms systems based on standard channel-robust acoustic features like MFCC-CMS and logRASTA-PLP. These frame-based front-ends are not effective for handling reverberation because they have been designed to reduce the effect of short-time spectral coloration. However, the duration of acoustic impulse responses is typically longer than the temporal analysis window (*e.g.*, 30ms). Therefore, the reverberation effect is no longer multiplicative in the short-term spectral domain and cannot be considered as spectral coloration.

4.3 Results on Real Speech Data

Our model selection method has been also tested on speech recorded in four different reverberant enclosures, namely the "office", "meeting", "lavatory" and "cafeteria" enclosures. The names of the enclosures are self-explanatory. For every reverberant environment, the anechoic test set is played back with a loudspeaker and simultaneously recorded with a distant omnidirectional microphone. The distance between the loudspeaker and the microphone is set to 2 meters. Room impulse responses are also measured using a correlation method based on an optimal time-stretched pulse (TSP) [44]. No attempt is made to compensate for the transfer functions of the loudspeaker and the microphone since both are studio-grade devices. The fullband reverberation times are estimated from Schroeder's decay curves [33] (*i.e.*, using the time-reverse integrals of the squared room impulse responses) and are reported in table 6.

First, we report cross-testing results (see table 7). The same conclusions as for simulated reverberated speech can be drawn: the best results are obtained when the model most closely matching the reverberation time is activated. Unfortunately, we observe that the recognition performance are not as good as for simulated reverberated speech when the training conditions match exactly the testing conditions, *e.g.*, for the "office" environment. Moreover, the performance degrade more severely if the conditions do not match. These differences can be explained by noting that the reverberation time T_{60} is not independent of the frequency while our reverberation process for training databases assumes so. Indeed, most real materials are more absorptive at high frequency and the absorption property of the air increases with frequency. Consequently, the reverberation time decreases with frequency in real acoustic environments leading to a mismatch with our proposed model. Nevertheless, our model selection method has been applied to the four real reverberated

test sets. The performance of the estimation algorithm are reported in table 8. Although the estimation algorithm does not perform as well as for simulated reverberated speech, the model selection method still yields significant improvements over the baseline ASR system (see table 9).

5 Summary and Concluding Remarks

In this communication, we have proposed a model selection method for automatic speech recognition in reverberant rooms. Our method consists in selecting the best acoustic model out of a library of models trained separately on artificially reverberated speech databases matching various reverberation times. The selection is based on the estimation of the fullband reverberation time. The estimation algorithm computes a Maximum Likelihood estimate of the reverberation time using only the short-term log-energy sequence of a reverberated speech utterance recorded in the reverberant operating environment

Experimental recognition tests on connected digit sequences have shown that the estimation algorithm performs well and the model selection approach improves significantly recognition accuracy for both simulated and real reverberated data. However, the performance improvement is not as high as expected on real data. This can be explained by the fact that the current modeling and estimation methods assume that the acoustic field is diffuse and that the reverberation time is independent of frequency. Further improvements can be expected by relaxing these severe assumptions. For example, we might consider estimating the reverberation time in frequency subbands. The training procedure should be modified accordingly.

Another shortcoming of the current model selection approach to hands-free speech recognition is that it deals only with reverberation: the speech signal is supposed to be noiseless. The proposed method should be extended to noisy environments to constitute a complete solution. This will be the subject of future work.

A Linear Distortion Model

Consider the distortion model of figure 2 and the LP-HMM of figure 4. Define $G(\cdot)$ as the transformation related the observation sequence Y_m to the LP-HMM emission sequence E_m given a state sequence s_m. The transformation $G(\cdot)$ is given by the following equations:

$$E_m = \sum_{p=0}^{1} \beta_p(s_m) X_{m-p}, \quad X_m = 10\log_{10}(W_m), \quad W_m = \sum_{k=0}^{1} \alpha_k Z_{m-k}, \quad Z_m = 10^{Y_m/10}. \qquad (22)$$

Clearly, the transformation is non-linear because of the logarithm and power functions. In section 3, we describe an iterative method for the ML estimation of the distortion parameters (α_0, α_1) based on the EM algorithm. In order to derive close-form re-estimation formulae, the transformation (22)

has to be linearized with respect to the distortion parameters. Let us assume that the successive estimates of the distortion parameters are varying slowly enough such that the sequence $W_m^{(\ell)} = \alpha_0^{(\ell)} Z_m + \alpha_1^{(\ell)} Z_{m-1}$ estimated at ℓ-th step is close to the next estimate $W_m^{(\ell+1)} = \alpha_0^{(\ell+1)} Z_m + \alpha_1^{(\ell+1)} Z_{m-1}$. Using a Taylor series expansion limited to the first order, we get the following linear approximation:

$$10\log_{10}(W_m^{(\ell+1)}) \simeq 10\log_{10}(W_m^{(\ell)}) + (W_m^{(\ell+1)} - W_m^{(\ell)})/\xi W_m^{(\ell)}, \quad \xi = \log(10)/10. \tag{23}$$

Equations (22) becomes

$$E_m^{(\ell+1)} \simeq \sum_{p=0}^{1} \beta_p(s_m) \left[\sum_{k=0}^{1} A_{k,m-p}^{(\ell+1)} Z_{m-p-k} + B_{m-p}^{(\ell+1)} \right] \tag{24}$$

with

$$A_{k,m}^{(\ell+1)} \triangleq \alpha_k^{(\ell+1)}/\xi W_m^{(\ell)} \tag{25}$$

and

$$B_m^{(\ell+1)} \triangleq X_m^{(\ell)} - 1/\xi. \tag{26}$$

If we neglect the boundary effects, that is, if we assume $Z_m = 0$ for $m < 0$ and $m > M$, equation (24) can be written in a matrix form for a finite-length sequence of observations $Z_{0,M} = \{Z_0, \ldots, Z_M\}$ and the corresponding finite-length sequence $E_{0,M}$ of emissions:

$$E_{0,M}^{(\ell+1)} = G(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha}^{(\ell+1)}) \simeq \mathbf{B}^{(\ell+1)} \left[\mathbf{A}^{(\ell+1)} Z_{0,M} + \mathbf{b}^{(\ell+1)} \right], \tag{27}$$

where the $(M+1) \times (M+1)$ matrices $\mathbf{B}$ and $\mathbf{A}$, and the $(M+1) \times 1$ vector $\mathbf{b}$ are defined by,

$$\mathbf{B}^{(\ell+1)} \triangleq \begin{bmatrix} \beta_0(s_0) & 0 & \ldots & 0 \\ \beta_1(s_1) & \beta_0(s_1) & & \vdots \\ 0 & \ddots & \ddots & \\ \vdots & & & 0 \\ 0 & \ldots & \beta_1(s_m) & \beta_0(s_m) \end{bmatrix}, \tag{28}$$

$$\mathbf{A}^{(\ell+1)} \triangleq \begin{bmatrix} A_{0,0}^{(\ell+1)} & 0 & \ldots & 0 \\ A_{1,1}^{(\ell+1)} & A_{0,1}^{(\ell+1)} & & \vdots \\ 0 & \ddots & \ddots & \\ \vdots & & & 0 \\ 0 & \ldots & A_{1,M}^{(\ell+1)} & A_{0,M}^{(\ell+1)} \end{bmatrix}, \tag{29}$$

and

$$\mathbf{b}^{(\ell+1)} \triangleq \begin{bmatrix} B_0^{(\ell+1)} \\ B_1^{(\ell+1)} \\ \vdots \\ B_M^{(\ell+1)} \end{bmatrix}, \tag{30}$$

respectively.

The approximation of the transformation $G(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha})$ is linear with respect to the distortion parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$. Using equations (22) and (27), its Jacobian matrix $\mathbf{J}$ can be computed as

$$\mathbf{J} \triangleq \frac{\partial G(Y_{0,M} \mid s_{0,M}, \boldsymbol{\alpha}^{(\ell+1)})}{\partial Y_{0,M}} = \frac{\partial G(\cdot)}{\partial Z_{0,M}} \cdot \frac{\partial Z_{0,M}}{\partial Y_{0,M}} = \mathbf{B}^{(\ell+1)} \mathbf{A}^{(\ell+1)} \mathbf{D}, \tag{31}$$

with $\mathbf{D}$ denoting a diagonal matrix whose elements are equal to $\xi Y_{0,M}$. It is straightforward to verify that the determinant of the Jacobian matrix $\mathbf{J}$ is given by

$$|\mathbf{J}| = |\mathbf{B}^{(\ell+1)}| \, |\mathbf{A}^{(\ell+1)}| \, |\mathbf{D}| = \prod_{m=0}^{M} \beta_0(s_m) A_{0,m}^{(\ell+1)} \xi Y_m. \tag{32}$$

References

[1] R. Siemund, H. Höge, S. Kunzmann and K. Marasek, "SPEECON – Speech Data for Consumer Devices", *Proc. of International Conference on Language Resources and Evaluation (LREC)*, vol. 2, pp. 883–886, Athens, Greece, May 2000.

[2] S. Nakamura and K. Shikano, "Room Acoustics and Reverberation: Impact on Hands-Free Recognition", *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 5, pp. 2419–2422, Rhodes, Greece, Sep. 1997.

[3] L. Couvreur, C. Couvreur and C. Ris, "A Corpus-Based Approach for Robust ASR in Reverberant Environments", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, vol. 1, pp. 397–400, Beijing, China, Oct. 2000.

[4] Y. Pan and A. Waibel, "The Effects of Room Acoustics on MFCC Speech Parameter", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, Oct. 2000.

[5] C. Avendano and H. Hermansky, "Study on the Dereverberation of Speech Based on Temporal Envelope Filtering", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, vol. 2, pp. 889–892, Philadelphia, USA, Oct. 1996.

[6] S. Subramaniam, A. P. Petropulu and C. Wendt, "Cepstrum-Based Deconvolution for Speech Dereverberation", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 5, pp. 392–396, Sep. 1996.

[7] D. Cole, M. Moody and S. Sridharan, "Position-Independent Enhancement of Reverberant Speech", *Journal of Audio Engineering Society*, vol. 45, no. 3, pp. 142–147, Mar. 1997.

[8] H. Nomura, S. Hirobayashi, T. Koike and M. Tohyama, "Dereverberation of Speech by Power Envelope Inverse Filtering", *Proc. of IEEE Workshop on Digital Signal Processing*, Bryce Canyon, USA, Aug. 1998.

[9] B. Yegnanarayana, P. M. Satyanarayanan, C. Avendano and H. Hermansky, "Enhancement of Reverberant Speech Using LP Residual", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 405–408, May. 1998.

[10] Q.-G. Liu, B. Champagne and P. Kabal, "A Microphone Array Processing Technique for Speech Enhancement in Reverberant Space", *Speech Communication*, vol. 18, no. 4, pp. 317–334, Jun. 1996.

[11] C. Marro, Y. Mahieux and K. U. Simmer, "Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Postfiltering", *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.

[12] F. Asano, S. Hayamizu, T. Yamada and S. Nakamura, "Speech Enhancement Based on the Subspace Method", *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 497–507, Sep. 2000.

[13] A. Koutras, E. Dermatas and G. Kokkinakis, "Improving Simultaneous Speech Recognition in Real Room Environments Using Overdetermined Blind Source Separation", *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 2, pp. 1009–1013, Aalborg, Denmark, Sep. 2001.

[14] R. Mukai, S. Araki and S. Makino, "Separation and Dereverberation Performance of Frequency Domain Blind Source Separation for Speech in a Reverberant Environment", *Proc. of European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 4, pp. 2599–2602, Aalborg, Denmark, Sep. 2001.

[15] B. D. Radlovič, R. C. Williamson and R. A. Kennedy, "Equalization in an Acoustic Reverberant Environment: Robustness Results", *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 311–319, May 2000.

[16] S. T. Neely and J. B. Allen, "Invertibility of a Room Impulse Response", *Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, Jan. 1979.

[17] M. Matassoni, M. Omologo and D. Giuliani, "Hands-Free Speech Recognition Using a Filtered Clean Corpus and Incremental HMM Adaptation", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1407–1410, Istanbul, Turkey, Jun. 2000.

[18] T. Takiguchi, S. Nakamura and K. Shikano, "HMM-Separation-Based Speech Recognition for a Distant Moving Speaker", *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, pp. 127–140, Feb. 2001.

[19] L. Couvreur, S. Dupont, C. Ris, J.-M. Boite and C. Couvreur, "Fast Adaptation for Robust Speech Recognition in Reverberant Environments", *Proc. of ISCA Workshop on Adaptation Methods For Automatic Speech Recognition*, Sophia Antipolis, France, Aug. 2001.

[20] L. Rigazio, D. Kryze, P. Nguyen and J.-C. Junqua, "Joint Environment and Speaker Adaptation", *Proc. of ISCA Workshop on Adaptation Methods For Automatic Speech Recognition*, Sophia Antipolis, France, Aug. 2001.

[21] K. Yamamoto, S. Nakagawa and H. Matsumoto, "Evaluation of PMC for Segmental Unit Input HMM in Various Environments" *Proc. of International Workshop on Hands-free Speech Communication*, pp. 183–186, Kyoto, Japan, Apr. 2001.

[22] Y. Zhao, "Statistical Estimation for Hands-Free Speech Recognition", *Proc. of International Workshop on Hands-free Speech Communication*, pp. 183–186, Kyoto, Japan, Apr. 2001.

[23] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[24] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[25] C. Avendano, S. Van Vuuren and H. Hermansky, "Data Based Filter Design for RASTA-like Channel Normalization in ASR", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, vol. 4, pp. 2087–2090, Philadelphia, USA, Oct. 1996.

[26] B. Kingsbury and N. Morgan, "Recognizing Reverberant Speech with RASTA-PLP", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 1259–1262, Munich, Germany, Apr. 1997.

[27] B. Kingsbury, N. Morgan and S. Greenberg, "Improving ASR Performance For Reverberant Speech", *Proc. of ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp. 87–90, Pont-à-Mousson, France, Apr. 1997.

[28] M. L. Shire and B. Y. Chen, "Data-Driven RASTA Filters in Reverberation", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 1627–1630, Istanbul, Turkey, Jun. 2000.

[29] M. L. Shire and B. Y. Chen, "On Data-derived Temporal Processing in Speech Feature Extraction", *Proc. of International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, Oct. 2000.

[30] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "Training of HMM with Filtered Speech Material for Hands-free Recognition", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 449–452, Phoenix, USA, Mar. 1999.

[31] V. Stahl, A. Fischer and R. Bippus, "Acoustic Synthesis of Training Data for Speech Recognition in Living Room Environments", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 21–24, Salt Lake City, USA, May 2001.

[32] M. Omura, M. Yada, H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Compensating of Room Acoustic Transfer Functions Affected by Change of Room Temperature", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 941–944, Phoenix, USA, Mar. 1999.

[33] H. Kuttruff, *Room Acoustics*, Elsevier, 4th ed., 2000.

[34] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.

[35] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[36] P. M. Peterson, "Simulating the Response of Multiple Microphones to a Single Acoustic Source in a Reverberant Room", *Journal of the Acoustical Society of America*, vol. 80, no. 5, pp. 1527–1529, Nov. 1986.

[37] J. Moorer, "About this Reverberation Business", *Computer Music Journal*, vol. 3, no. 2, pp. 13–28, 1979.

[38] C. J. Wellekens, "Explicit Time Correlation in Hidden Markov Models for Speech Recognition", *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 384–387, Dallas, USA, Apr. 1987.

[39] P. Kenny, M. Lennig and P. Mermelstein, "A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 2, pp. 220–225, Feb. 1990.

[40] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, ser. B, vol. 39, pp. 1–38, 1997.

[41] AURORA database - http://www.elda.fr/aurora2.html.

[42] A. Papoulis, "Probability, Random Variables, and Stochastic Processes", McGraw-Hill, 3rd ed., 1991.

[43] H. Bourlard and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Publishers, 1994.

[44] Y. Suzuki, F. Asano, H.-Y. Kim and T. Sone, "An Optimum Computer-Generated Pulse Signal Suitable for the Measurement of Very Long Impulse Responses", *Journal of the Acoustical Society America*, vol. 97(2), pp. 1119–1123, Feb. 1995.

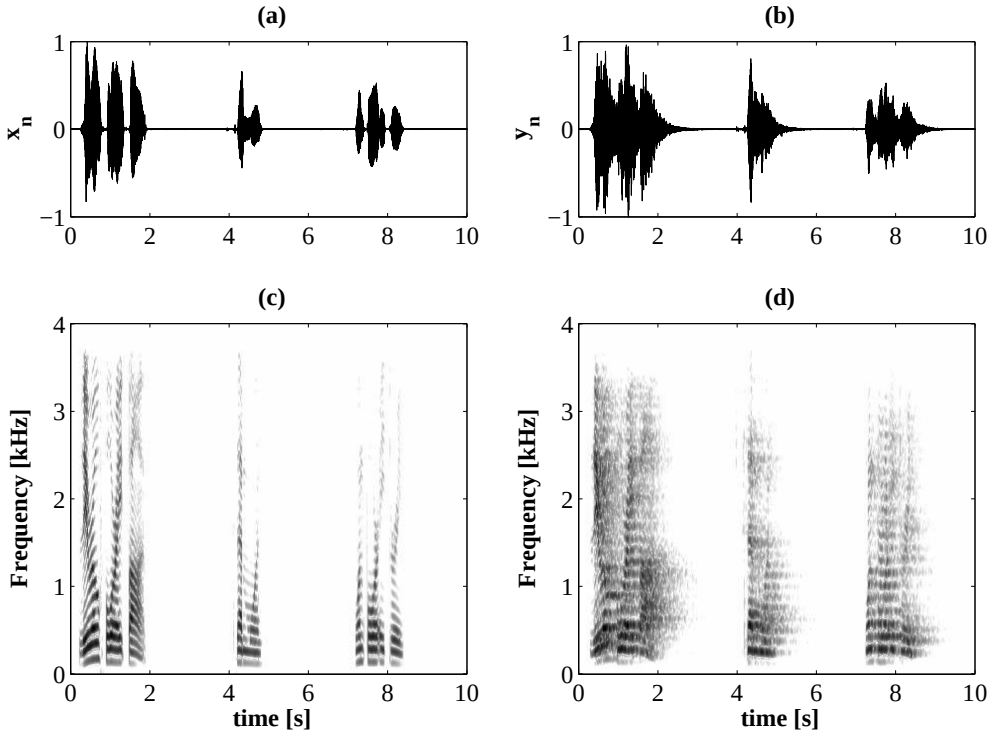

Figure 1: Waveforms of (a) an anechoic speech utterance x_n (digit sequence "z485-71-1483") and (b) its reverberant version y_n; (c) and (d) show the respective spectrograms.

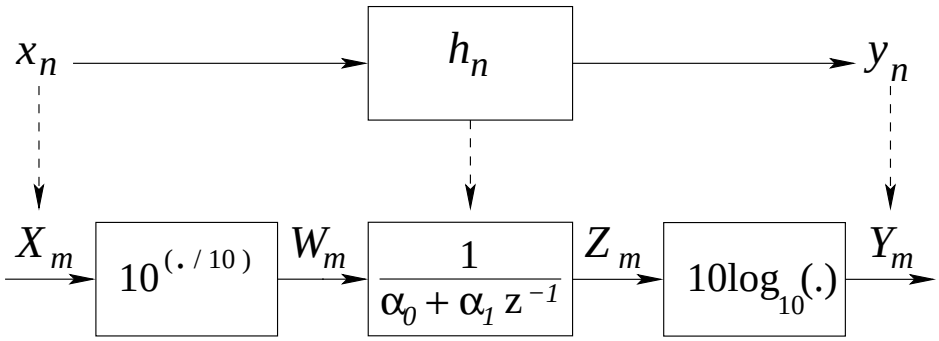

Figure 2: Room reverberation model (upper part) for temporal signals and equivalent diffuse model (lower part) for short-term log-energy sequences.

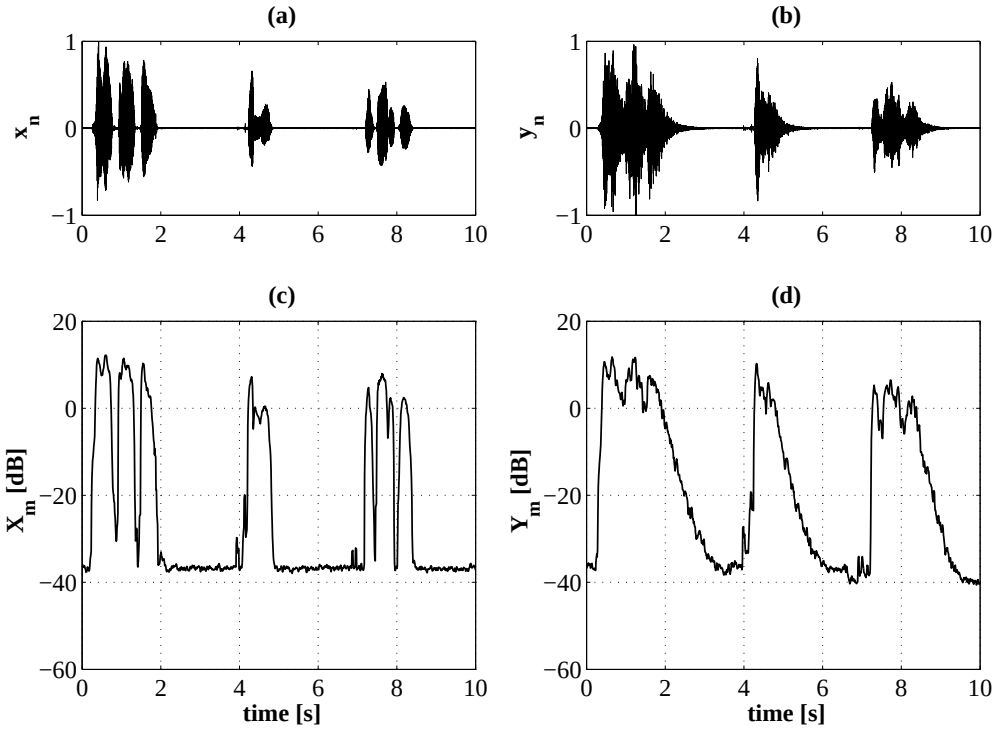

Figure 3: Waveforms of (a) an anechoic speech utterance x_n (digit sequence "z485-71-1483") and (b) its reverberant version y_n; (c) and (d) show the respective short-term log-energy sequences X_m and Y_m for T_w=30ms and F_r=100Hz.

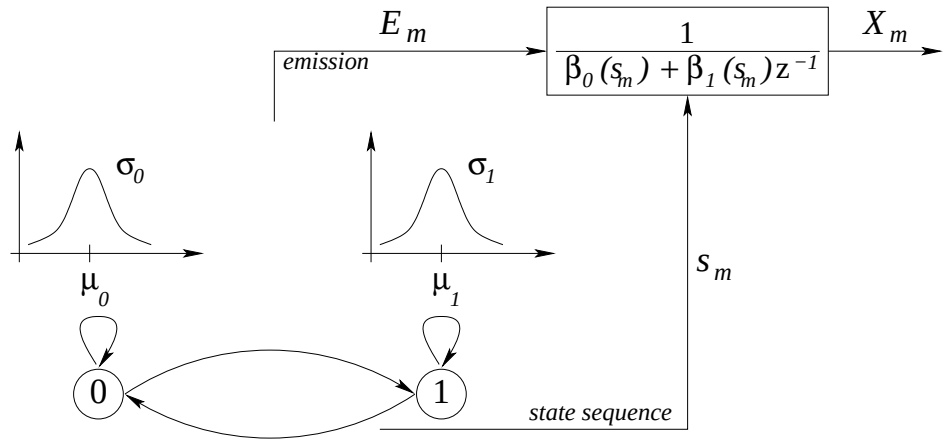

Figure 4: 2-state one-dimensional first-order LP-HMM for modeling short-term log-energy sequences of anechoic speech (0: silence state, 1: speech state).

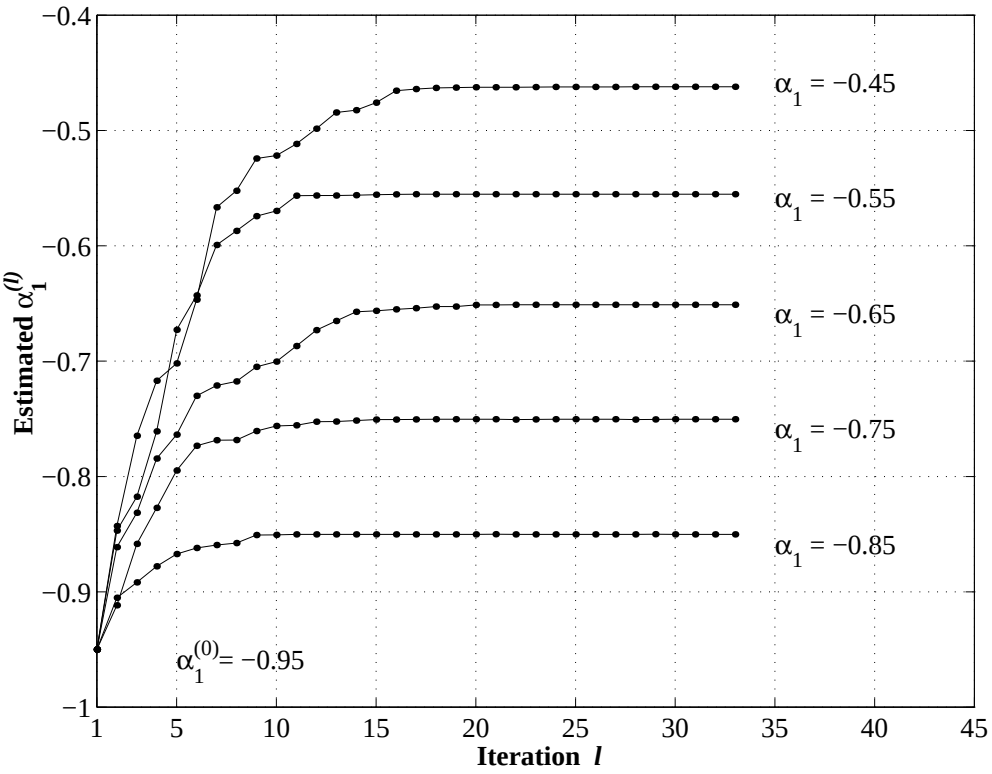

Figure 5: Convergence plot of the estimation algorithm for various values of the distortion parameter α_1. The length $M + 1$ of the observation sequence is set to 256.

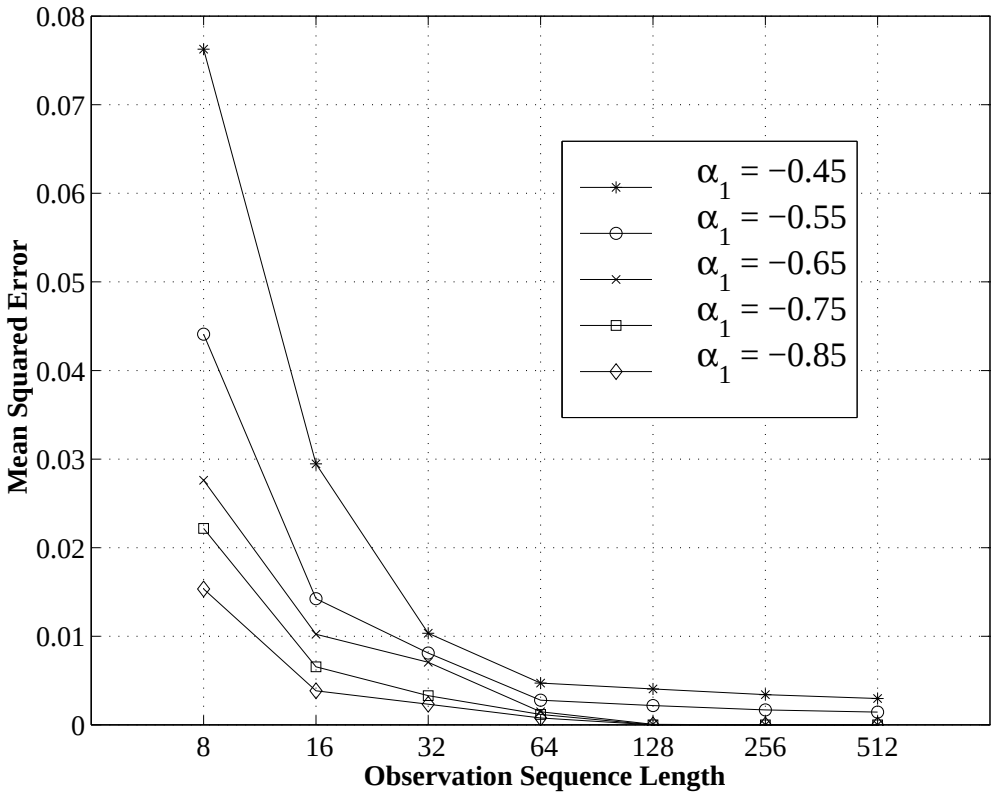

Figure 6: Mean squared error (MSE) for the distortion parameter α_1 as a function of the observation sequence length $M + 1$ and for various true values of the distortion parameter.

Table 1: Parameters of a 2-state one-dimensional first-order LP-HMM for modeling short-term log-energy sequences of anechoic speech (T_w=30ms and F_r=100Hz).

$s_m = i$	a_{ii}	a_{ij}	μ_i	σ_i	$\beta_0(i)$	$\beta_1(i)$
0	0.95	0.05	-4.3	4.2	1.0	-0.92
1	0.03	0.97	1.1	3.2	1.0	-0.77

Table 2: Iterative algorithm for the estimation of the fullband reverberation time T_{60} from the short-term log-energy sequence $Y_{0,M}$ of a reverberated speech utterance.

1. Initialize the estimate of the distortion parameter $\alpha_1^{(0)}$ ($\alpha_0 = 1$); set $\ell = 0$; and compute $Z_m = 10^{Y_m/10}$, $m = 0, \ldots, M$.

2. Apply the current inverse distortion filter to obtain $W_m^{(\ell)} = \alpha_0^{(\ell)} Z_m + \alpha_1^{(\ell)} Z_{m-1}$; and compute $X_m^{(\ell)} = 10 \log_{10} W_m^{(\ell)}$, $m = 0, \ldots, M$.

3. Estimate the *a posteriori* state probabilities $\gamma_{m,i}^{(\ell)}$ via the Forward-Backward algorithm [39] given the LP-HMM parameters and $X_m^{(\ell)}$, $m = 0, \ldots, M$.

4. Update the estimate of $\alpha_1^{(\ell+1)}$ by applying the re-estimation formula (17)–(21).

5. Set $\ell = \ell + 1$ and go to 2 unless convergence is reached.

6. Derive the reverberation time T_{60} from $\alpha_1^{(\ell)}$ via (5).

Table 3: Performance (WER [%]) of MFCC-based acoustic models trained on artificially reverberated speech for various simulated reverberant testing conditions.

Test set	Training set								
	anechoic	$T_{60} = 200$ms	400	600	800	1000	1200	1400	1600
anechoic	**1.7**	2.9	7.6	11.9	15.9	19.8	20.6	22.7	23.8
$T_{60} = 200$ms	7.0	**3.6**	4.5	6.4	9.8	12.5	13.7	15.1	16.1
300	7.8	**3.9**	**4.4**	6.4	9.8	12.3	13.9	15.0	15.7
400	18.7	9.6	**5.2**	5.7	8.5	12.2	12.8	14.7	15.3
500	20.1	11.2	**5.9**	**5.9**	8.7	12.2	12.8	14.7	15.4
600	29.7	20.2	11.3	**9.2**	10.0	12.6	13.7	15.6	16.4
700	33.2	24.7	14.9	**11.2**	**11.3**	13.6	14.4	16.1	17.1
800	41.0	33.7	22.1	17.3	**14.0**	15.9	16.6	18.2	19.1
1000	43.4	35.8	24.0	20.4	16.0	**17.0**	17.1	18.7	19.7
1200	49.3	43.1	32.0	27.9	20.9	20.7	**20.4**	21.6	22.1
1400	51.1	48.5	36.8	33.5	26.0	24.9	23.2	**24.5**	24.4
1600	52.9	50.1	37.3	36.6	28.1	26.7	25.3	25.1	**25.1**

Table 4: Median absolute value error (MAVE) [ms] and relative MAVE [%] for blind estimation of T_{60} and corresponding confusion rate (CR) [%] for model selection in the case of simulated reverberated speech.

Test set	MAVE [ms]	relative MAVE [%]	CR [%]
200ms	53.4	29.2	3.3
300ms	37.8	12.6	1.2
400ms	67.9	17.0	10.8
500ms	64.1	12.8	9.0
600ms	70.8	11.8	22.7
800ms	88.7	11.1	10.7
1000ms	52.4	5.2	20.6
1200ms	95.0	7.9	47.8
1400ms	103.8	7.4	51.1
1600ms	130.3	8.1	31.8
Average	76.4	12.3	20.9

Table 5: Comparison of the performances (WER (SUB/DEL/INS) [%]) of the baseline system, two standard normalization techniques, the T_{60}-based model selection method and the "Oracle" method for simulated reverberated speech. Scores are averaged over all testing environments.

Method	WER [%]	SUB/DEL/INS [%]
MFCC-baseline	32.2	9.9 / 18.6 / 3.7
MFCC-CMS	37.5	8.7 / 27.4 / 1.4
logRASTA-PLP	36.1	11.0 / 24.1 / 1.0
T_{60}-based mod. sel.	13.3	4.7 / 5.9 / 2.7
"Oracle" mod. sel.	12.7	4.6 / 5.2 / 2.9

Table 6: Fullband reverberation times T_{60} [ms] measured by Schroeder's method in real reverberant environments.

Room	Description	T_{60} [ms]
office	medium size room with windows and concrete floor and walls	604.0
meeting	medium size room with glass windows and cement floor	873.3
lavatory	small room with tiled floor and walls	1307.1
cafeteria	large room with wooden floor and glass walls	1461.8

Table 7: Performances WER [%] of MFCC-based acoustic models trained on artificially reverberated speech for various real reverberant testing conditions.

Test set	Training set								
	anechoic	$T_{60} = 200$ms	400	600	800	1000	1200	1400	1600
anechoic	**1.7**	2.9	7.6	11.9	15.9	19.8	20.6	22.7	23.8
office	30.2	24.8	13.8	**11.2**	15.6	19.0	21.2	22.7	27.2
meeting	47.7	44.8	29.2	23.0	**20.5**	23.7	28.0	28.0	30.4
lavatory	57.9	59.2	50.5	47.9	43.7	39.6	36.9	**33.9**	37.9
cafeteria	57.8	56.4	47.9	42.7	38.4	35.6	36.1	**32.9**	37.7

Table 8: Median absolute value error (MAVE) [ms] and relative MAVE [%] for blind estimation of T_{60} and corresponding confusion rate (CR) [%] for model selection in the case of real reverberated speech.

Test set	MAVE [ms]	relative MAVE [%]	CR [%]
office	143.8	23.8	34.0
meeting	120.8	13.8	21.8
lavatory	126.1	9.7	19.9
cafeteria	137.4	9.4	22.8
Average	132.0	14.2	24.6

Table 9: Comparison of the performances WER (SUB/DEL/INS) [%] of the baseline system, two standard normalization techniques, the T_{60}-based model selection method and the "Oracle" method for real reverberated speech. Scores are averaged over all testing environments.

Method	WER [%]	SUB/DEL/INS [%]
MFCC-baseline	48.4	22.3 / 21.8 / 4.2
MFCC-CMS	53.3	14.6 / 36.9 / 1.8
logRASTA-PLP	51.9	19.9 / 27.2 / 4.8
T_{60}-based mod. sel.	28.2	14.7 / 8.8 / 4.7
"Oracle" mod. sel.	24.7	12.8 / 7.5 / 4.4