Distant-talking Continuous Speech Recognition based on a novel Reverberation Model in the Feature Domain

Armin Sehr, Marcus Zeller, and Walter Kellermann

Multimedia Communications and Signal Processing,
University of Erlangen-Nuremberg
Cauerstr. 7, 91058 Erlangen, Germany
{sehr,zeller,wk}@LNT.de

Abstract

A novel approach for automatic speech recognition in highly reverberant environments, proposed in [1] for isolated word recognition, is extended to continuous speech recognition (CSR) in this paper. The approach is based on a combined acoustic model consisting of a network of clean speech HMMs and a reverberation model. Because the grammatical information and the information about the acoustic environment are strictly separated in the combined model, a high degree of flexibility for adapting the system to new tasks and new environments is attained. We show that virtually all known CSR search algorithms can be used for decoding the proposed combined model if a few extensions are added. In a simulation of a connected digit recognition task, the proposed method achieves more than 40 % reduction of the word error rate compared to a conventional HMM-based system trained on reverberant speech, at the cost of an increased decoding complexity.

Index Terms: robust speech recognition, distant-talking speech recognition, dereverberation.

1. Introduction

Automatic speech recognition (ASR) is the key to numerous applications like natural human-machine interfaces, dictation systems, electronic translators and automatic information desks. To further increase the acceptance of these applications, it is desirable that the user can move freely while communicating to the system without the need of wearing a close-talking microphone.

Since the distance between speaker and microphone in such a distant-talking scenario usually is in the range of one to several meters, unwanted additive signals and reverberation of the desired signal hamper ASR. In this paper, we focus on reverberation-robust ASR.

The most straightforward approach of obtaining an ASR system capable of working in reverberant environments is to train a conventional HMM-based recognizer using data recorded in the enclosure where the recognizer will be deployed. To reduce the enormous effort implied in collecting a complete set of training data for each new environment of operation, artificial reverberation of clean training data has been suggested [2, 3] and has been shown to yield a noticeable improvement.

While the usual model adaptation techniques, which have been successfully applied in noisy environments, are not suitable for reverberation times significantly exceeding the frame length of the recognizer, Raut et al. [4] suggest a model adaptation approach for long reverberation. Here, the linear means of a split-state HMM are adjusted taking into account the linear means of the preceding states. Thus, the recognition rate is significantly improved with only a few adaptation data. However, both reverberant training and model adaptation techniques suffer from the underlying assumption of 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.

In this paper, we extend the novel approach for robust speech recognition in reverberant environments introduced in [1] to continuous speech recognition. The dependence of the current feature vector on previous vectors is implicitly accounted for in this approach by a combined acoustic model consisting of a network of conventional HMMs, modeling the clean speech, and a reverberation model. Since the HMM network is independent of the acoustic environment, it needs to be trained only once using the usual Baum-Welch re-estimation procedure. The training of the reverberation model is based on a set of room impulse responses for the corresponding acoustic environment and involves only a negligible computational effort. In this way, the recognizer can be adapted to new environments with moderate effort.

The paper is organized as follows: In Section 2, the novel approach is explained in detail. Simulations of a connected digit recognition task, described in Section 3, show the effectiveness of the new recognizer. In Section 4, the paper is summarized and conclusions are drawn.

2. The proposed approach

The combined acoustic model is introduced from the perspective of feature production and its novel part, namely the reverberation model, is described in detail. We show that virtually all known CSR search algorithms can be used for decoding the combined acoustic model if the calculation of the feature vector output probabilities is adapted accordingly.

2.1. Feature production model

We assume that the sequence $\mathbf{X}$ of reverberant speech feature vectors $\mathbf{x}(n)$ is produced by a combined acoustic model. The combined model consists of a network $\mathcal{N}_\lambda$ of word-level HMMs λ_p describing the clean speech and a reverberation model η as illustrated in Figure 1. The word-level HMMs λ_p may be composed of subword HMMs. The task grammar and the language model can be embedded into the network of HMMs to reflect the recognition task. In contrast to that, the reverberation model is completely independent of the recognition task. The strict separation of the grammatical information incorporated into the network of HMMs

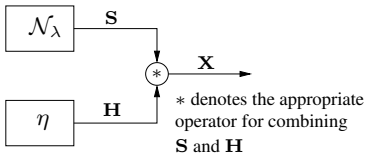

Figure 1: Proposed feature production model.

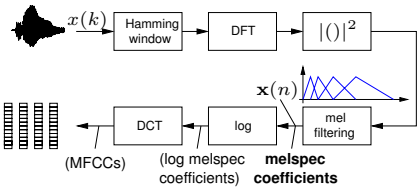

Figure 2: Calculation of melspec coefficients.

and the information about the acoustic environment reflected by the reverberation model yields a high degree of flexibility when the recognition system has to be adapted to new tasks or new acoustic environments. By selecting an appropriate reverberation model, the combined model can be used in moderately to highly reverberant environments and also with non-reverberant speech.

The combined acoustic model for the production of reverberant feature vectors can be applied to any kind of speech features which allow the formulation of an appropriate relation between the sequence $\mathbf{S}$ of output feature vectors $\mathbf{s}(n)$ of the clean speech model, the sequence $\mathbf{H}$ of the reverberation model output matrices $\mathbf{H}(n)$ (see 2.2) and the sequence $\mathbf{X}$ of reverberant speech feature vectors $\mathbf{x}(n)$.

In this paper, we are using mel-frequency spectral (melspec) coefficients as illustrated in Figure 2 as features. Thus, the reverberant sequence $\mathbf{X}$ can be approximated by the convolution of the clean sequence $\mathbf{S}$ and the sequence $\mathbf{H}$ of realizations of the reverberation model

$$\mathbf{x}(n) = \sum_{m=0}^{M-1} \mathbf{h}(m,n) \odot \mathbf{s}(n-m) \quad \forall\, n = 1 \ldots N+M-1\,. \quad (1)$$

Here, $\odot$ denotes element-wise multiplication, $\mathbf{s}(n)$ and $\mathbf{x}(n)$ are single feature vectors at frame index n of clean and reverberant speech, respectively, the vector $\mathbf{h}(m,n)$ is a realization of the reverberation model for frame delay m and frame index n, while M and N are the lengths of the reverberation model and the clean utterance, respectively.

Note that the proposed combined model can be considered as a generalization of the HMM decomposition approach [5] for a convolutive combination of the models employing a different way of evaluating the output density of the combined model.

2.2. Reverberation model

The reverberation model η represents an independent identically distributed (iid) matrix-valued random process. 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. The sequence $\mathbf{H}$ of reverberation feature matrices $\mathbf{H}(n)$ is a realization of this random process as illustrated in Figure 3. For simplification, each element of the matrix is assumed to be statistically independent from all other elements and is modeled by a Gaussian density. Furthermore, the iid property of the random process implies that all elements of the random process at frame index n_1 are statistically independent from all elements of the random process at frame index n_2 as long as $n_1 \neq n_2$.

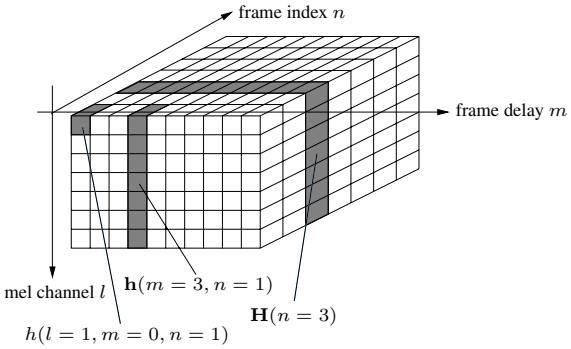

Figure 3: Realization $\mathbf{H}$ of the reverberation model η.

The starting point for the training of the reverberation model is a set of room impulse responses (RIRs) for different microphone and loudspeaker positions of the room where the ASR system will be applied. These RIRs can either be measured before using the recognizer, estimated by blind system identification approaches or modeled, e. g., using the image method as described in [6]. To train the reverberation model, the RIRs are time-aligned so that the direct path of all RIRs appears at the same delay. Calculation of the melspec representation yields a matrix of melspec coefficients for each impulse response. Using these coefficients, the means and the variances of all matrix elements of η are estimated.

2.3. Decoding

So far, we introduced a novel feature production model, describing how reverberant speech features are generated given the model. For speech recognition however, the opposite task has to be solved. Given a reverberant utterance, a recognition network of clean speech HMMs and a reverberation model, the recognizer has to find the path through the network yielding the highest probability for the utterance in connection with the reverberation model.

Independently of the acoustic-phonetic modeling, the continuous speech recognition search problem can be formulated as finding the word sequence $\hat{W}$ maximizing the product of the language model score $L(W)$ associated with word sequence W and the acoustic model score $A(\mathbf{X}|W)$ of $\mathbf{X}$ given W

$$\hat{W} = \underset{W}{\operatorname{argmax}} \left\{ L(W) \cdot A(\mathbf{X}|W) \right\}\,. \quad (2)$$

If conventional HMMs are used for the acoustic-phonetic modeling, the acoustic score can be expressed as the following maximum likelihood problem

$$A(\mathbf{X}|W) = \max_{Q} \left\{ P(\mathbf{X}, Q|\Lambda) \right\}\,,$$

where the maximization is performed over all allowed state sequences Q through the sequence Λ of HMMs describing W.

For the combined acoustic model consisting of a clean speech HMM network and the reverberation model, the acoustic score is given as

$$\begin{aligned} A(\mathbf{X}|W) &= \max_{Q,\mathbf{S},\mathbf{H}} \left\{ P(Q, \mathbf{S}, \mathbf{H}|\Lambda, \eta) \right\} \quad \text{s. t. (1)} \\ &= \max_{Q} \left\{ P(Q|\Lambda) \cdot \max_{\mathbf{S},\mathbf{H}} \left\{ P(\mathbf{S}, \mathbf{H}|\Lambda, \eta, Q) \right\} \right\} \\ &\quad \text{subject to (s. t.) the constraint (1)}\,. \end{aligned}$$

A large number of algorithms exists for solving the search problem (2) if HMMs are used for the acoustic modeling, see [7]

and [8] for overviews. As only the calculation of the acoustic score is different in the proposed approach, virtually all these algorithms can be used for decoding the combined model if a few extensions are added. These extensions will be derived in the following.

In the proposed approach, the acoustic score $A(\mathbf{X}|W)$ is calculated iteratively by an extended version of the Viterbi algorithm

$$\begin{aligned}
\gamma_j(n) &= \max_i\{\gamma_i(n-1)\cdot a_{ij}\cdot O_{ij}(n)\}, \\
&\quad \forall j=1\ldots I, \quad n=2\ldots N+M-1, \\
O_{ij}(n) &= \max_{\mathbf{s}_{ij}(n),\mathbf{H}_{ij}(n)}\{\, f_\Lambda(j,\mathbf{s}_{ij}(n))\cdot f_\eta(\mathbf{H}_{ij}(n))\} \quad (3) \\
\text{s. t.} \quad \mathbf{x}(n) &= \sum_{m=0}^{M-1}\mathbf{h}_{ij}(m,n)\odot\mathbf{s}_{ij}(n-m)\;, \quad (4) \\
A(X|W) &= \gamma_I(N+M-1)\;.
\end{aligned}$$

Here, $\gamma_j(n)$ is the Viterbi metric for state j at frame n, a_{ij} is the transition probability from state i to state j, $f_\Lambda(j,\mathbf{s}_{ij}(n))$ and $f_\eta(\mathbf{H}_{ij}(n))$ are the output densities of the HMM sequence Λ describing W and the reverberation model η, respectively, I is the number of states in Λ. The subscript ij in $\mathbf{s}_{ij}(n)$ and $\mathbf{H}_{ij}(n)$ indicates that these vectors/matrices are based on the optimum partial state sequence $\hat{Q}_{ij}(n)$ from frame $n-M+1$ to frame n with current state j and previous state i.

The extension compared to the conventional Viterbi algorithm consists of the inner optimization of equation (3). Introducing a simplified notation which neglects the dependency on the frame index n and the sequence $\hat{Q}_{ij}(n)$ by the following mappings $\mathbf{s}_{ij}(n-m)\rightarrow\mathbf{s}_m$, $\mathbf{x}(n)\rightarrow\mathbf{x}$, $\mathbf{h}_{ij}(m,n)\rightarrow\mathbf{h}_m$, the constraint (4) can be written as

$$\mathbf{x}=\underline{\mathbf{h}_0}\odot\underline{\mathbf{s}_0}+\sum_{m=1}^{M-1}\underline{\mathbf{h}_m}\odot\overline{\mathbf{s}_m}\;,$$

where the *underlined vectors* are unknowns following a Gaussian distribution with diagonal covariance matrix and the *overlined vectors* are known from previous steps of the algorithm.

Now we approximate the generally non-Gaussian random vector $\tilde{\mathbf{x}}_0=\mathbf{h}_0\odot\mathbf{s}_0$ resulting from the element-wise product of the two Gaussian random vectors $\mathbf{h}_0$ and $\mathbf{s}_0$ by a Gaussian random vector $\mathbf{x}_0$ with the same mean and variance as $\tilde{\mathbf{x}}_0$. Thus we can rewrite the constraint as

$$\mathbf{x}=\underline{\mathbf{x}_0}+\sum_{m=1}^{M-1}\underline{\mathbf{h}_m}\odot\overline{\mathbf{s}_m}\;. \quad (5)$$

A two-step closed-form solution of the constrained problem (3) s. t. (4) can be derived in the following way.

First step: Find $\mathbf{x}_0$ and $\mathbf{h}_{m'}$.
Applying the method of Lagrange multipliers to

$$\max_{\mathbf{x}_0,\mathbf{h}_1,\ldots,\mathbf{h}_{M-1}}\{\, f_{\mathbf{x_0}}(\mathbf{x}_0)\cdot f_\eta(\mathbf{h}_1)\cdot\ldots\cdot f_\eta(\mathbf{h}_{M-1})\}\quad\text{s. t. (5)}\;,$$

where $f_{\mathbf{x_0}}(\mathbf{x}_0)$ is the probability density of $\mathbf{x}_0$ and assuming that all involved densities are single Gaussians, we obtain the following solutions for $\mathbf{x}_0$ and $\mathbf{h}_{m'}$

$$\begin{aligned}
\mathbf{x}_0 &= \frac{\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}{\sigma^2_{\mathbf{x}_0}+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}\odot m_{\mathbf{x}_0} \\
&+ \frac{\sigma^2_{\mathbf{x}_0}}{\sigma^2_{\mathbf{x}_0}+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}\odot\Big(\mathbf{x}-\sum_{m=1}^{M-1}\mathbf{s}_m\odot m_{\mathbf{h}_m}\Big)
\end{aligned}$$

$$\begin{aligned}
\mathbf{h}_{m'} &= \frac{\sigma^2_{\mathbf{x}_0}+\sum_{\substack{m=1\\m\neq m'}}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}{\sigma^2_{\mathbf{x}_0}+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}\odot m_{\mathbf{h}_{m'}} \\
&+ \frac{\mathbf{s}^2_{m'}\odot\sigma^2_{\mathbf{h}_{m'}}}{\sigma^2_{\mathbf{x}_0}+\sum_{m=1}^{M-1}\mathbf{s}_m^2\odot\sigma^2_{\mathbf{h}_m}}\odot\frac{1}{\mathbf{s}_{m'}}\odot\Big(x-m_{\mathbf{x}_0}-\sum_{\substack{m=1\\m\neq m'}}^{M-1}\mathbf{s}_m\odot m_{\mathbf{h}_m}\Big)
\end{aligned}$$

where the squaring and the division operation denote element-wise squaring and element-wise division, respectively, and $m_{\mathbf{h}_m}$ and $\sigma^2_{\mathbf{h}_m}$ denote the mean and the variance vector of $\mathbf{h}_m$, respectively, and likewise for the other variables.

Second step: Find $\mathbf{h}_0$ and $\mathbf{s}_0$ given $\mathbf{x}_0$.
Applying the method of Lagrange multipliers to

$$\max_{\mathbf{s}_0,\mathbf{h}_0}\{\, f_\Lambda(j,\mathbf{s}_0)\cdot f_\eta(\mathbf{h}_0)\}\quad\text{s. t.}\quad\overline{\mathbf{x}_0}=\underline{\mathbf{h}_0}\odot\underline{\mathbf{s}_0}$$

and assuming that $f_\Lambda(j,\mathbf{s}_0)$ and $f_\eta(\mathbf{h}_0)$ are single Gaussians, we obtain the following fourth-order equation to be fulfilled by the desired vector $\mathbf{h}_0$

$$\sigma^2_{\mathbf{s}_0}\odot\mathbf{h}_0^4-m_{\mathbf{h}_0}\odot\sigma^2_{\mathbf{s}_0}\odot\mathbf{h}_0^3+m_{\mathbf{s}_0}\odot\sigma^2_{\mathbf{h}_0}\odot\mathbf{x}_0\odot\mathbf{h}_0-\mathbf{x}_0^2\odot\sigma^2_{\mathbf{h}_0}=0\,,$$

where the exponents denote element-wise powers. It can be shown, that this equation has a pair of complex conjugate solutions, one real-valued positive and one real-valued negative solution. As only the real-valued positive solution achieves the maximization of the desired probability, we obtain exactly one vector $\mathbf{h}_0$ and thus exactly one vector $\mathbf{s}_0$.

Note that for this two-step solution of the inner optimization problem, the vectors $\mathbf{s}_{ij}(n-m)\;\forall\;m=1\ldots M-1$ need to be stored for all possible previous states i and all possible current states j in the HMM sequence Λ. This can be implemented effectively by storing all speech vectors $\mathbf{s}(n-m|q(n-m)=k)$ for all possible states k in Λ. Then $\mathbf{s}_{ij}(n-m)$ can be reconstructed using the optimum partial state sequence $\hat{Q}_{ij}(n)$. Therefore, the used search algorithm needs to perform state-level backtracking so that $\hat{Q}_{ij}(n)$ is available.

With these extensions, the known CSR search algorithms can be used for decoding the proposed combined model.

3. Simulations

To analyze the effectiveness of the proposed approach, simulations of a connected digit recognition task using melspec features are carried out. The performance of the proposed approach is compared to that of conventional HMM-based recognizers (using the same melspec features) trained on clean and reverberant speech, respectively.

3.1. Experimental setup

The proposed approach is implemented by extending the functionality of HTK [9] with the inner optimization as described above. For the evaluation of the proposed approach, a connected digit recognition task is chosen, since this can be considered as one of the easiest examples of continuous speech recognition.

The used feature vectors are calculated in the following way: The speech signal, sampled at 20 kHz, is decomposed into overlapping frames of length 25 ms with a frame shift of 10 ms. After applying a 1st-order pre-emphasis (coefficient 0.97) and a Hamming window, a 512-point DFT is computed. From the DFT representation, 24 melspec coefficients are calculated. Only static features and no Δ and $\Delta\Delta$ coefficients are used.

The training is performed using 4579 connected digit utterances corresponding to 1.5 hours of speech from the TI digits [10]

training data. For the training with reverberant speech, the clean data are convolved with measured room impulse responses from two different rooms. Room A is a lab environment with a reverberation time of $T_{60} = 300\,\text{ms}$ and a signal-to-reverberation ratio of $SRR = 4\,\text{dB}$. Room B is a studio environment with $T_{60} = 700\,\text{ms}$ and $SRR = -4\,\text{dB}$.

A 16-state left-to-right model without skips over states is trained for each of the 11 digits ('0'-'9' and 'oh'). Additionally, a three-state silence model with a backward skip from state 3 to state 1 is trained. The output densities are single Gaussians with diagonal covariance matrices. All HMMs are trained according to the following procedure: First, single Gaussian MFCC-based HMMs are trained by 10 iterations of Baum-Welch re-estimation. Then the melspec HMMs are obtained from the MFCC HMMs by single pass retraining [11]. In this way, more reliable models are obtained than by training melspec models from scratch. For the conventional HMM-based clean recognizer and for the proposed approach, identical HMM networks are used. The HMM network of the conventional reverberant recognizer differs only with respect to the training data. Two distinct sets of reverberant HMMs are trained for room A and room B using data reverberated with RIRs measured in the corresponding rooms.

We use HMMs with single Gaussian densities for the proposed approach, as the solution to the inner optimization problem described above is only valid for single Gaussians. To get an equal comparison, single Gaussian HMMs are also used for the conventional approaches.

For the recognition, a silence model is added in the beginning and at the end of the HMM network consisting of an 11-digit loop. As test data, 512 test utterances randomly selected from the TI digits test set are used. To obtain the reverberant feature sequences, the clean test signals are convolved with room impulse responses from room A and room B, respectively, before they are passed to the feature extraction unit.

To train the reverberation model η_A/η_B for room A/B with length $M_A = 20/M_B = 50$, 36/18 impulse responses measured in room A/B with different loudspeaker and microphone positions with constant distance of 2.00 m/4.12 m are used. For the artificial reverberation of training data and for the training of the reverberation models, RIRs different from the impulse responses used to generate the test data (measured in the same room but at different microphone positions) are used in order to maintain a strict separation of training and test data.

3.2. Experimental results

Table 1 compares the word error rates (WER) of the conventional HMM-based recognizers to that of the proposed approach for the connected digit recognition task described above. While the WER increase in room A compared to clean speech is about 30 % and about 15 % for the conventional systems trained on clean and reverberant speech, respectively, the error rate of the proposed approach only increases by less than 5 %. In the more reverberant room B, the benefit of the proposed approach is even greater with a WER increase of about 9 % compared to about 70 % and 27 % of the conventional systems. These results confirm that the proposed approach achieves much better recognition performance in reverberant environments than conventional ASR systems, even if the latter are trained on reverberant data. However, the decoding complexity increases by a multiplicative factor which is proportional to the length M of the reverberation model. In our current implementation, this factor is in the range of one thousand.

Test	conventional clean training	conventional rev. training	proposed
clean data	17.98 %	-	-
rev. data - room A	48.50 %	33.17 %	22.39 %
rev. data - room B	87.08 %	45.14 %	26.36 %

Table 1: Comparison of word error rates of a conventional HMM-based recognizer and of the proposed algorithm.

4. Summary and conclusions

A novel approach for continuous speech recognition in reverberant environments has been presented. The method uses a combination of an HMM network and a reverberation model to describe the reverberant speech feature sequences. As the HMM network modeling the clean speech is identical to the networks used in conventional CSR, virtually all search algorithms developed for conventional CSR can be used for decoding the combined acoustic models. The search algorithms only need to be extended by an inner optimization procedure accounting for the reverberation model. The limitations of this concept are the increased computational complexity and the memory requirements for this inner optimization. Simulations of a connected digit recognition task have shown a considerably better performance of the proposed approach compared to conventional HMM-based recognizers, even if the latter are trained on reverberant speech. Future work will focus on implementing the proposed method for more powerful speech features, like mel-frequency cepstral coefficients, and on using mixtures of Gaussians as well as on reducing the computational complexity.

5. References

[1] A. Sehr, M. Zeller, and W. Kellermann, "Hands-free speech recognition using a reverberation model in the feature domain," *Proc. European Signal Processing Conference (EUSIPCO)*, September 2006, to appear.

[2] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of HMM with filtered speech material for hands-free recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 449–452, March 1999.

[3] V. Stahl, A. Fischer, and R. Bippus, "Acoustic synthesis of training data for speech recognition in living-room environments," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 285–288, May 2001.

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

[5] A. P. Varga and R. K. Moore, "Hidden markov model decomposition of speech and noise," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 845–848, 1990.

[6] 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, April 1979.

[7] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice Hall, Upper Saddle River, NJ, USA, 2001.

[8] H. Ney and S. Orthmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 64–63, September 1999.

[9] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2)*, Cambridge University Engineering Department, Cambridge, UK, 2002.

[10] R. G. Leonard, "A database for speaker-independent digit recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 42.11.1–42.11.4, 1984.

[11] P. C. Woodland, M. J. F. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 65–68, May 1996.