A NEW CONCEPT FOR FEATURE-DOMAIN DEREVERBERATION FOR ROBUST DISTANT-TALKING ASR

Armin Sehr and Walter Kellermann

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

ABSTRACT

The feature-domain dereverberation capabilities of a novel approach for automatic speech recognition in reverberant environments are investigated in this paper. By combining a network of clean speech HMMs and a reverberation model, the most likely combination of the HMM output and the reverberation model output is found during decoding time by an extended version of the Viterbi algorithm. We show in this paper that the most likely HMM output represents a good estimate of the clean speech feature sequence and can be used as input to subsequent speech recognizers.

Index Terms— robust speech recognition, distant-talking speech recognition, dereverberation, feature-domain processing, Viterbi decoding

1. INTRODUCTION

Automatic speech recognition (ASR) is widely used in applications like human-machine interfaces, dictation systems, electronic translators and automatic information desks. For many of these applications, hands-free operation is desirable, i. e. the user can talk from a distance and move freely while communicating to the system without the need of wearing a close-talking microphone.

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

One possibility to achieve reverberation robustness is to dereverberate the speech signal before it is processed by the recognizer. Blind dereverberation is an extremely challenging task, since neither the room impulse response (RIR) describing the acoustic path between speaker and microphone nor the speaker signal are available. In [1] Yegnanarayana et al. propose to attenuate additional impulses caused by reverberation in the linear prediction residual signal.

Using two or more microphones, exact inverse filtering is possible [2]. Direct estimation of multi-channel inverse filters is suggested in [3] exploiting three fundamental properties of speech signals, namely nonwhiteness, nongaussianity, and nonstationarity. In [4] the harmonic structure of speech is used for inverse filter determination.

For robust speech recognition, we do not need to estimate the clean speech waveform. It is sufficient to estimate a clean speech feature sequence. As the feature sequence contains less information than the waveform, we believe that finding the clean speech feature sequence is simpler than estimating the dereverberated waveform.

We propose to use a novel approach for robust speech recognition in reverberant environments, first introduced in [5] and [6], as a preprocessing step to estimate clean speech feature sequences from reverberant feature sequences directly in the feature domain. The approach is based on combining a network of clean speech HMMs and a reverberation model. In the decoding phase, the most likely combination of the HMM output and the reverberation model output is found by maximum likelihood estimation. We show in this paper that the most likely HMM output represents an estimate of the clean speech feature sequence and can be used as input to subsequent speech recognizers.

The paper is organized as follows: In Section 2, the approach introduced in [5, 6] is reviewed. Its application for feature-domain dereverberation is then discussed in Section 3. Section 4 evaluates the dereverberation capabilities by simulations and Section 5 concludes the paper.

2. THE REVERBERATION MODEL BASED SPEECH RECOGNITION APPROACH

We assume that the sequence $\mathbf{X}$ of reverberant speech feature vectors $\mathbf{x}(n)$ is produced by a combination of a network $\mathcal{N}_\lambda$ of HMMs λ_p describing the clean speech and a reverberation model η as illustrated in Figure 1.

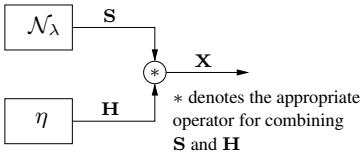

Fig. 1. Proposed feature production model.

If linear mel-frequency spectral (melspec) coefficients are used as features, as assumed throughout the paper, 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, and the vector $\mathbf{h}(m,n)$ is a realization of the reverberation model for frame lag m and frame index n, while M and N are the lengths of the reverberation model and the clean utterance, respectively.

The reverberation model η is a statistical representation of the acoustic path between speaker and microphone in the feature domain and can be considered as an iid matrix-valued random process (see [6]).

Independently of the acoustic-phonetic modeling, the 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

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

For the combined acoustic model according to Figure 1, the acoustic score is given as

$$\begin{aligned} A(\mathbf{X}|W) &= \max_{Q,\mathbf{S},\mathbf{H}} \{P(Q,\mathbf{S},\mathbf{H}|\Lambda,\eta)\} \quad \text{s. t. (1)} \\ &= \max_{Q} \left\{ P(Q|\Lambda) \cdot \max_{\mathbf{S},\mathbf{H}} \{P(\mathbf{S},\mathbf{H}|\Lambda,\eta,Q)\} \right\} \\ &\quad \text{subject to (s. t.) the constraint (1)} \ . \end{aligned} \tag{3}$$

where Λ is the sequence of HMMs modeling W and Q is a possible sequence of states through Λ.

If the acoustic score calculation accounts for the reverberation model as given in Equation (3), most of the known search algorithms (see e. g. [7], chapter 13 or [8]) can be used to solve the problem (2).

Accounting for the reverberation model is achieved by calculating the acoustic score $A(\mathbf{X}|W)$ iteratively by an extended version of the Viterbi algorithm as given by

$$\gamma_j(n) = \max_i \{\gamma_i(n-1) \cdot a_{ij} \cdot O_{ij}(n)\}, \quad \forall j = 1 \dots I, \quad n = 2 \dots N+M-1, \tag{4}$$

$$O_{ij}(n) = \max_{\mathbf{s}_{ij}(n), \mathbf{H}_{ij}(n)} \{ f_\Lambda(j, \mathbf{s}_{ij}(n)) \cdot f_\eta(\mathbf{H}_{ij}(n)) \} \tag{5}$$

$$\text{s. t.} \quad \mathbf{x}(n) = \sum_{m=0}^{M-1} \mathbf{h}_{ij}(m,n) \odot \mathbf{s}_{ij}(n-m) \ , \tag{6}$$

$$A(\mathbf{X}|W) = \gamma_I(N+M-1) \ .$$

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 correspond to the current state j and previous state i. Details about solving the inner optimization problem (5), representing the key extension, are given in [6].

In this way, the combined acoustic model can be decoded, given the reverberant feature sequence $\mathbf{X}$, in order to find the most likely word sequence $\hat{W}$. While in [6], this approach is used directly to find the best transcription $\hat{W}$, we propose in the following section to use this approach as a preprocessing unit to perform dereverberation in the feature domain.

3. FEATURE DOMAIN DEREVERBERATION

In the following, we propose a feature-domain dereverberation algorithm based on the extended Viterbi decoding, which determines an estimate $\hat{\mathbf{S}}$ of the clean speech feature sequence $\mathbf{S}$ corresponding to $\mathbf{X}$.

Figure 2 shows a schematic overview of the extended Viterbi recursion. The inner optimization is solved in two steps. First, the estimated clean speech feature vector $\mathbf{s}_{ij}(n)$ and the estimated feature-domain RIR matrix $\mathbf{H}_{ij}(n)$ are determined according to the constrained optimization problem (5, 6). To calculate $O_{ij}(n)$, these values are inserted into the corresponding densities. The vector $\mathbf{s}_{ij}(n)$ is the most likely clean speech estimate for frame n, current state j and predecessor state i.

Once the maximization (4) over all possible predecessor states i has been performed, the most likely clean speech estimate $\mathbf{s}_j(n)$ for frame n and state j can be selected from the $\mathbf{s}_{ij}(n)$ vectors in the following way

$$\begin{aligned} \hat{i} &= \underset{i}{\operatorname{argmax}} \{\gamma_i(n-1) \cdot a_{ij} \cdot O_{ij}(n)\} \ , \\ \mathbf{s}_j(n) &= \mathbf{s}_{\hat{i}j}(n) \ . \end{aligned}$$

For each state j and each frame n, $\mathbf{s}_j(n)$ is stored in a matrix of clean speech vectors (see Figure 2).

When the entire utterance is decoded, the most likely state sequence $\hat{Q} = \hat{q}(1) \dots \hat{q}(N+M-1)$ corresponding to $\mathbf{X}$ can be determined by conventional Viterbi backtracking (see e. g. [7], chapter 8). Using this state sequence, the sequence of most likely clean speech vectors $\hat{\mathbf{S}} = \hat{\mathbf{s}}(1) \dots \hat{\mathbf{s}}(N+M-1)$ can be extracted from the matrix of clean speech vectors as given by

$$\hat{\mathbf{s}}(n) = \mathbf{s}_{j=\hat{q}(n)}(n) \ .$$

The sequence $\hat{\mathbf{S}}$ is the most likely output of the clean speech HMM network, given the reverberant sequence $\mathbf{X}$, the HMM network $\mathcal{N}_\lambda$ and the reverberation model η. Thus it can be considered as a dereverberated version of $\mathbf{X}$.

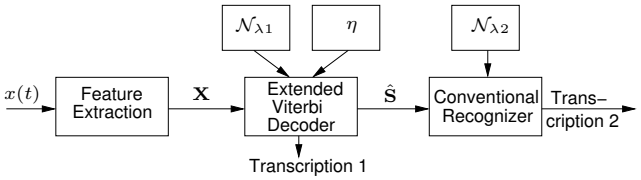

Fig. 3. Extended Viterbi algorithm as preprocessing unit.

The dereverberated feature sequence $\hat{\mathbf{S}}$ can be used as input to a subsequent speech recognizer as shown in Figure 3. The HMM network $\mathcal{N}_{\lambda 1}$ used by the extended Viterbi decoder and the HMM network $\mathcal{N}_{\lambda 2}$ may be different. E. g. the output densities of $\mathcal{N}_{\lambda 1}$ may be single Gaussian densities while the output densities of $\mathcal{N}_{\lambda 2}$ may be mixtures of Gaussians.

In this way, the extended Viterbi decoder can be used as preprocessing unit performing dereverberation in the feature domain.

4. SIMULATIONS

To analyze the dereverberation capabilities of the proposed approach, simulations of a connected digit recognition task using melspec features are carried out. First we compare the melspec representations of a clean, a reverberant and a dereverberated utterance. Then the recognition rate of four different approaches is compared: conventional recognizers trained on clean and reverberant speech, respectively, applied to reverberant speech data, the approach of [6] and the approach proposed in Section 3.

4.1. Experimental setup

The functionality of HTK [9] is extended by the approach of [6] and by an estimation of the clean speech sequence $\hat{\mathbf{S}}$ as proposed in Section 3. Connected digit recognition is chosen as a simple example of continuous speech recognition.

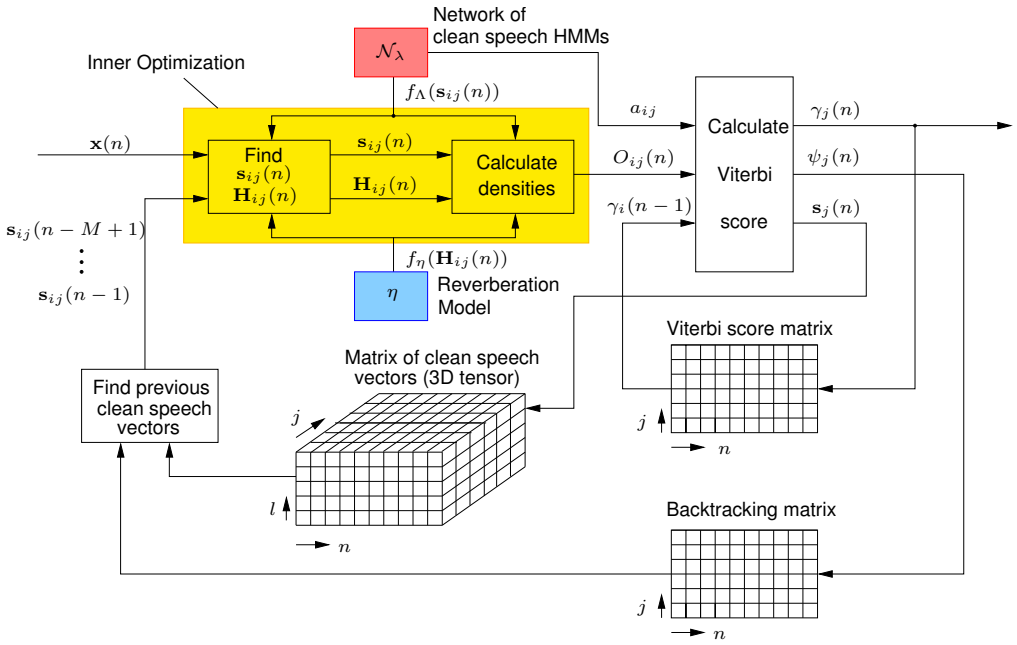

Fig. 2. Illustration of the extended Viterbi recursion.

The static feature vectors are calculated in the following way: 24 melspec coefficients are calculated from the speech signal, sampled at 20 kHz and 1st-order pre-emphasized with a coefficient of 0.97, using a 512-point DFT based on a Hamming window of length 25 ms and a frame shift of 10 ms, no Δ and $\Delta\Delta$ coefficients are used.

The training is performed using 4579 connected digit utterances from the TI digits [10] training data. For the training with reverberant speech, the clean data are convolved with measured RIRs 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'). Both single Gaussian output densities and mixtures of three Gaussians are used.

For the recognition, a silence model is added at 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 RIRs 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 (see [6]). For the artificial reverberation of training data and for the training of the reverberation models, RIRs different from the RIRs 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.

4.2. Experimental results

Figure 4 shows the melspec representation of the utterance "four, two, seven" using a dB color scale. Comparing the clean (close talk recording, Figure 4 a)) and the reverberant utterance (room B, Figure 4 b)), the dispersive effect of reverberation on the feature vector sequence is clearly visible. E. g. the short period of silence before the plosive /t/ in "two" or the low energy region of the lower mel channels for the fricative /s/ in "seven" is largely filled with energy from the preceding vowels in the reverberant utterance.

In the feature sequence generated according to the approach proposed in Section 3, (Figure 4 c)) these regions of low energy are restored to a large extent and the dispersion across frames is clearly reduced. The rough-textured fashion of the dereverberated utterance results from the assumption of statistical independence between different channels and different frames.

Word accuracies in %		room A		room B	
		number of Gaussians			
recognizer	input data	1	3	1	3
I conv. clean training	$\mathbf{X}$	51.5	63.7	13.4	14.0
II conv. reverb. training	$\mathbf{X}$	66.8	80.4	54.6	72.1
III approach of [6]	$\mathbf{X}$	77.6	-	71.6	-
IV conv. clean training	$\hat{\mathbf{S}}$	77.5	80.5	71.7	72.1

Table 1. Comparison of word accuracies of a conventional HMM-based recognizer, trained on clean and reverberant speech, the approach of [6] and the feature-domain dereverberation proposed in Section 3 for single Gaussians and mixtures of three Gaussians.

Table 1 shows the word accuracies for conventional recognizers trained on clean speech (I) and reverberant speech (II), respectively, applied to the reverberant test sequences $\mathbf{X}$, the approach proposed in [6] (III) and the approach proposed in Section 3 (IV). For methods I and II, the results are given both for single Gaussian densities and for mixtures of three Gaussians. As approach III is currently only implemented for single Gaussian densities, no results for three Gaussian mixtures can be reported. The results of IV are based on the clean speech feature estimates based on single Gaussian densities ($\mathcal{N}_{\lambda 1}$ in Figure 3 uses single Gaussian output densities). For the conventional recognizer working on the dereverberated features, both single Gaussian densities and mixtures of three Gaussians are

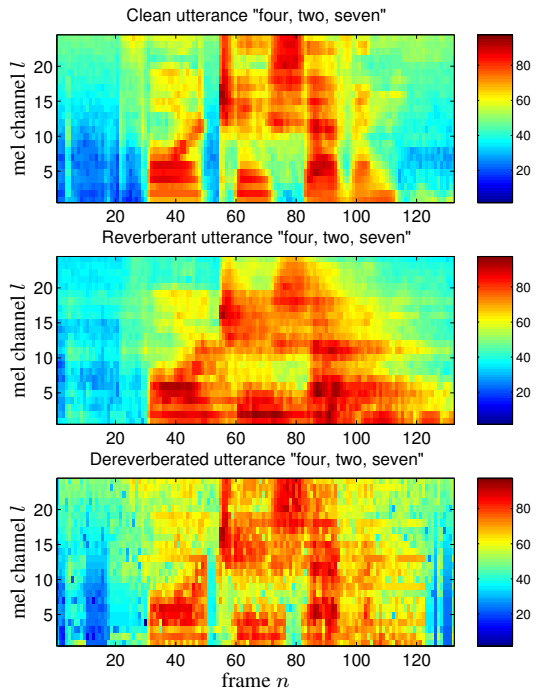

Fig. 4. Comparison of a) clean feature sequence $\mathbf{S}$, b) reverberant feature sequence $\mathbf{X}$, and c) dereverberated feature sequence $\hat{\mathbf{S}}$ in the melspec domain using a dB color scale.

used ($\mathcal{N}_{\lambda 2}$ in Figure 3).

Table 1 confirms the findings of [6], namely that the approach III is significantly more robust to reverberation than the conventional approaches I and II using single Gaussian densities. In room B, approach III is even close to the reverberantly trained HMMs with mixtures of three Gaussians.

Using the estimated clean feature sequence in a conventional recognizer (IV) with single Gaussian output densities results in an almost identical word accuracy as obtained by approach III. Indeed, the transcriptions of both methods are virtually identical. This can be expected because the estimated clean speech feature sequence is biased towards the initial transcription (Transcription 1 in Figure 3). Therefore, IV produces the same errors as III if identical clean speech models are used.

If mixtures of three Gaussian are used in the output densities of $\mathcal{N}_{\lambda 2}$, some of the errors of the initial transcription can be corrected and the performance is increased. While a clear increase in performance is observed in room A (from 77.6 % to 80.5 %), only a marginal increase is obtained in room B (from 71.6 % to 72.1 %).

These results confirm the dereverberation capability of the approach IV. It is indeed remarkable, that the final conventional recognizer trained on clean speech can achieve similar results as the extremely reverberation robust approach III using the same clean speech models. Using different clean speech models $\mathcal{N}_{\lambda 1}$ and $\mathcal{N}_{\lambda 2}$ the performance of IV can be increased. However, further investigation is required to optimize the combination of the clean speech models $\mathcal{N}_{\lambda 1}$ and $\mathcal{N}_{\lambda 2}$ in order increase the gain in performance of IV compared to III.

Future work includes evaluating the use of a clean speech model $\mathcal{N}_{\lambda 1}$ with increased variance or a phoneme-class-based HMM network $\mathcal{N}_{\lambda 1}$ instead of a network of word-level HMMs to decrease the dependence of $\hat{\mathbf{S}}$ on the initial transcription and thus to increase the possibility of correcting errors of the initial transcription in the second pass. Optimization of the balance between the clean speech model $\mathcal{N}_{\lambda 1}$ and the reverberation model η in the extended Viterbi algorithm will also be analyzed.

If a greater gain can be achieved by the final recognizer, a three-pass version of the feature-domain dereverberation approach IV will become attractive: In a first pass, the state/frame-alignment is determined by a conventional recognizer. In the second pass, the feature domain dereverberation is performed based on the inner optimization of Equations (5) and (6). Because then this optimization has to be performed only once for each frame, the computational complexity is significantly reduced compared to the full extended Viterbi search. In the third pass, the dereverberated feature sequence is used in a conventional recognizer to find the final transcription.

5. SUMMARY AND CONCLUSIONS

We showed in this paper that the reverberation model based speech recognition method proposed in [6] can be extended to a feature-domain dereverberation approach, if the most likely clean speech estimate is extracted and used as input to a conventional recognizer for clean speech. A comparison of a reverberant utterance and the corresponding estimated clean speech feature sequence showed the feature-domain dereverberation capability of the proposed approach.

Applying a conventional recognizer to the dereverberated feature sequence achieves a similar performance as the approach of [6] if the same single Gaussian clean speech models are used. Using a clean speech model with mixtures of three Gaussians, a slightly increased performance could be achieved.

Future work includes the optimization of the variations between the clean speech models $\mathcal{N}_{\lambda 1}$ and $\mathcal{N}_{\lambda 2}$ in order to increase the gain of the second pass and the implementation of a three-pass ASR system based on feature-domain dereverberation with reduced computational complexity.

6. REFERENCES

[1] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 267–281, May 2000.

[2] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, February 1988.

[3] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: A versatile framework for multichannel blind signal processing," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. III, pp. 889–892, May 2004.

[4] T. Nakatani, M. Miyoshi, and K. Kinoshita, "Implementation and effects of single channel dereverberation based on the harmonic structure of speech," *Proc. IEEE Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 91–94, September 2003.

[5] 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.

[6] A. Sehr, M. Zeller, and W. Kellermann, "Distant-talking continuous speech recognition based on a novel reverberation model in the feature domain," *Proc. International Conference on Spoken Language processing (ICSLP/INTERSPEECH)*, September 2006.

[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.