SCALED FACTORIAL HIDDEN MARKOV MODELS: A NEW TECHNIQUE FOR COMPENSATING GAIN DIFFERENCES IN MODEL-BASED SINGLE CHANNEL SPEECH SEPARATION

M. H. Radfar\textsuperscript{1}, W. Wong\textsuperscript{1}, R. M. Dansereau\textsuperscript{2}, and W.-Y. Chan\textsuperscript{3}

\textsuperscript{1}Department of Electrical and Computer Engineering, University of Toronto, Canada
\textsuperscript{2}Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada
\textsuperscript{3}Department of Electrical and Computer Engineering, Queen’s University, Kingston, Canada

\section*{ABSTRACT}

In model-based single channel speech separation, factorial hidden Markov models (FHMM) have been successfully applied to model the mixture signal $Y(t) = X(t) + V(t)$ in terms of trained patterns of the speech signals $X(t)$ and $V(t)$. Nonetheless, when the test signals are scaled versions of the trained patterns (i.e. $g_x X(t)$ and $g_v V(t)$), the performance of FHMM degrades significantly. In this paper, we introduce a modification to FHMM, called scaled FHMM, which compensates gain difference. In this technique, first, the scale factors are expressed in terms of the target-to-interference ratio (TIR). Then, an iteration quadratic optimization approach is coupled with FHMM to estimate TIR which with the decoded HMM sequences maximize the likelihood of the mixture signal. Experimental results, conducted on 180 mixtures with TIRs from 0 to 15 dB, show that the proposed technique significantly outperforms unscaled FHMM, and scaled/unscaled vector quantization speech separation techniques.

\textbf{Index Terms}— source separation, factorial hidden Markov models, model-based single channel speech separation, quadratic optimization, and mixmax approximation.

\section{1. INTRODUCTION}

Separating the speech signals $X(t)$ and $V(t)$ from the mixture $Y(t) = X(t) + V(t)$—with no spatial diversity—using speakers’ pre-trained models is commonly referred to as model-based single channel speech separation (SCSS). Factorial hidden Markov models (FHMM) \cite{1} was first applied to this problem by Rowies \cite{2} whose technique inspired succeeding works in multi-talker speech recognition and separation in the single channel paradigm \cite{3–5}. In FHMM-based SCSS, in the training phase, the feature space (log spectral vectors) of each speaker is modeled using an HMM. Then, in the separation phase, the state sequences of two HMMs which maximize the likelihood of the mixture are decoded using the parallel Viterbi algorithm \cite{6}. As a drawback, FHMM-based SCSS only model the log-spectral patterns assuming that the data used in the training and separation phases have identical loudness, i.e. $Y(t) = g_x X(t) + g_v V(t)$, $g_x = g_v$. In practice, however, we encounter situations where $g_x \neq g_v$. Therefore, it is of great importance to develop techniques which handle the gain mismatch problem. Gain adaptation were already addressed in the previous works \cite{7, and references therein} where other modeling approaches such as vector quantization (VQ) were used. In this paper, we extend the solution to FHMM in which separation does not take place at the frame level. In this method, $g_x$ and $g_v$ are first expressed in terms of the target-to-interference ratio $\theta$ (scale factor) which is included as an unknown parameter into FHMM. Having the gain-included FHMM, the task is now to not only decode the best state sequences but also the scale factor. This is performed using a quadratic optimization approach which is augmented to FHMM. We show through experiments that the proposed scaled FHMM significantly outperforms the unscaled FHMM and VQ techniques.

\section{2. PRELIMINARY DEFINITIONS AND MODELS}

Let $X(t)$ and $V(t)$, $t = 0, \ldots, T - 1$, be the target and interference speech signals. Let $g_x$ and $g_v$ be two positive real values which represent the scale factors. The observation signal $Y(t)$ is then given by

$$Y(t) = g_x X(t) + g_v V(t), \quad t = 0, \ldots, T - 1. \quad (1)$$

It is assumed that the signals have equal power before gain scaling.

In line with the previous techniques, we use the log spectral vectors of the windowed speech files as the input feature. Therefore, here we present the notations used for representing log spectral vectors of the observation, target, and interference signals. Let $Y(t)$, $X(t)$, and $V(t)$ be split into $R$ overlapping frames. The log spectral vectors corresponding to the observation, target and interference for the $r^{th}$ frame are given, respectively, by $y^r = \{y^r(d)\}_{d=0}^{D-1}$, $x^r = \{x^r(d)\}_{d=0}^{D-1}$, and $v^r = \{v^r(d)\}_{d=0}^{D-1}, r = 1, \ldots, R$. The relation between $y^r$, $x^r$ and $v^r$ can be expressed using the MIXMAX approximation \cite{6}. According to the MIXMAX approximation, the log spectrum of the observation is almost exactly equal to the maximum element-wise of the log spectra of the target and interference. Mathematically, this means

$$y^r(d) \approx \max\{g(\theta) + x^r(d), g(-\theta) + v^r(d)\} \quad d = 0, \ldots, D - 1 \quad (3)$$

where $\log_{10} g_x = g(\theta)$ and $\log_{10} g_v = g(-\theta)$ and $A = \frac{g_x^2}{g_v^2}$, $g(x)$ is the gain function. It is clear that estimating $g_x$ and $g_v$ is equivalent to estimating $\theta$ since $A$ is known in advance. Therefore, we, hereafter, focus on estimating $\theta$.

We denote the $d$-th frame as $f_d$ and $|$ denotes the length of the frame.

The authors wish to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding this project.
3. SCALED FHMM

A detailed description of FHMM (also known as parallel HMMs) in context of model-based SCSS was given in [6]. In the training phase, there is no difference between FHMM and scaled FHMM where we simply fit an HMM to the feature space of each speaker. There is however a modification in the decoding (separation) phase where the two HMMs are combined. Consistent with the definition given in [6], let \( \lambda_i(x^s, a^s, b^s) \) and \( \lambda^*(a^*, b^*) \) represent the HMM parameter sets for \( x^s \) and \( v^s \), respectively. HMM parameters comprise the initial probabilities \( \pi \), the transition probabilities \( a \), and the observation probabilities \( b \). Having these models, we aim at finding the best HMMs state sequences \( \hat{q}_x = (\hat{q}_{i1}, \hat{q}_{i2}, \ldots, \hat{q}_{ik}) \), and \( \hat{q}_v = (\hat{q}_{i1}, \hat{q}_{i2}, \ldots, \hat{q}_{jk}) \), which maximize

\[
\hat{q}_x, \hat{q}_v = \arg\max_{q_x, q_v} p(q_x, q_v, y | \lambda_x, \lambda_y, \theta).
\] (4)

For an arbitrary \( \theta \), the maximization problem in (4) is solved using the parallel Viterbi algorithm [6] in the five steps:

1. Preprocessing
   - \( \hat{\pi}_x = \log \pi_x \), and \( \hat{\pi}_v = \log \pi_v \), \( 1 \leq j, k \leq K \)
   - \( \hat{b}_{i,j,k}(y^r | \theta) = \log b_{i,j,k}(y^r | \theta) \), \( 1 \leq j, k \leq K \), \( 1 \leq r \leq R \)
   - \( \hat{a}_{i,j} = \log a_{i,j} \), and \( \hat{a}_{i,k} = \log a_{i,k} \) \( 1 \leq i, j, k \leq K \)

2. Initialization
   - \( \hat{\delta}_1(j, k, \theta) = \log \delta_1(j, k, \theta) = \hat{\pi}_x + \hat{\pi}_v + \hat{b}_{i,j,k}(y^r | \theta) \) \( 1 \leq j, k \leq K \)
   - \( \psi_0(j, k) = 0 \) \( 1 \leq j, k \leq K \), \( 1 \leq r \leq R \)

3. Recursion
   - \( \hat{\delta}_r(j, k, \theta) = \log \delta_r(j, k, \theta) = \max_{1 \leq i, \ell \leq K} \{ \hat{\delta}_{r-1}(i, \ell, \theta) + \hat{a}_{i,j} + \hat{a}_{i,k} + \hat{b}_{i,j,k}(y^r | \theta) \} \) \( 1 \leq j, k \leq K \), \( 2 \leq r \leq R \)
   - \( \psi_r(j, k) = \arg\max_{1 \leq i, \ell \leq K} \{ \hat{\delta}_{r-1}(i, \ell, \theta) + \hat{a}_{i,j} + \hat{a}_{i,k} \} \) \( 1 \leq j, k \leq K \), \( 2 \leq r \leq R \)

4. Termination
   - \( P(\hat{q}_x^0, \hat{q}_v^0 | \theta) = \max_{1 \leq i, \ell \leq K} \hat{\delta}_R(i, \ell, \theta) \)
   - \( (\hat{q}_x^0, \hat{q}_v^0) = \arg\max_{1 \leq i, \ell \leq K} \hat{\delta}_R(i, \ell, \theta) \)

5. Path backtracking
   - \( (\hat{q}_x^r, \hat{q}_v^r) = \psi_{r+1}(\hat{q}_x^{r+1}(\theta), \hat{q}_v^{r+1}(\theta)) \) \( r = R - 1, R - 2, \ldots, 1 \)

Including \( \theta \) in the parallel Viterbi algorithm only affects the observation probability \( b_{i,j,k}(y^r | \theta) \). In [8, IV. A], we obtained an approximation to \( p(y^r | d) = q_x^r = j, q_v^r = k, \theta \) in terms of the PDFs of \( x^r \) and \( v^r \) when \( g_x = g_v = 0 \). Adding \( \log_{10} g_x = g(\theta) \) and \( \log_{10} g_v = g(-\theta) \) to \( x^r \) and \( v^r \) shifts only the means of the PDFs of \( x^r \) and \( v^r \) by \( g(\theta) \) and \( g(-\theta) \), respectively. Thus, \( b_{i,j,k}(y^r | \theta) = \end{align*}

\[
- \sum_{d=0}^{D-1} \frac{1}{2} \left( g^r(d) - \frac{1}{2} \log \sigma_{\max}(d) \right)^2
- \log \sigma_{\max}(d) - \frac{1}{2} \log 2\pi
\] (5)

4. ESTIMATING \( \theta \) USING QUADRATIC OPTIMIZATION

In this section, we propose an approach to estimate the scale factor \( \theta \). Suppose that \( \theta \) lies in the interval \( \Theta = [\theta_{\min}, \theta_{\max}] \) over which the separation of two signals is feasible. This means that for those \( \theta \) outside the range of \( \Theta \), one source is significantly dominant such that \( g_x \gg g_v \rightarrow Y(t) \approx g_x X(t) \) or vice versa. In this paper, we set \( \theta_{\min} = -15 \text{ dB} \) and \( \theta_{\max} = 15 \text{ dB} \). For those \( \theta \) outside of this interval, the weaker signal is almost completely masked by the stronger signal such that the separation is practically impossible.

The likelihood function obtained for the pairs of best state sequences \( \hat{q}_x^v \) and \( \hat{q}_v^v \) is a function of \( \theta \), that is, \( Q(\theta) = P(\hat{q}_x^v, \hat{q}_v^v ) \).

This means that for different values of \( \theta \), different pairs of sequences and likelihoods are obtained. We are, however, interested to find

\[
\hat{\theta} = \arg\max_{\theta} Q(\theta).
\] (6)

By examining the curves of \( Q(\theta) \) versus \( \theta \) for several examples, we found that \( Q(\theta) \) is approximated well by a quadratic function (a polynomial of degree 2) whose maximum corresponds to \( Q(\hat{\theta}) \).

Therefore, we can apply an iterative quadratic optimization approach to find \( \hat{\theta} \). In the experiments, we observed that the algorithm for finding \( \hat{\theta} \) converges after only two or three iterations using quadratic optimization.

Fig. 1 shows a block diagram of the proposed method for decoding the paths \( \{\hat{q}_x^v, \hat{q}_v^v \} \) and \( \theta \). We assume that \( \theta \) lies in the interval \([-15 \text{ dB}; 15 \text{ dB}] \) and then the paths corresponding to \( \theta \) are decoded using the parallel Viterbi algorithm. Using the decoded paths and the quadratic optimization approach, a new value for \( \theta \) is obtained.

The algorithm is iterated until either a global maximum is found or until a maximum number of iterations is reached. The detailed procedure for performing the quadratic optimization is described in [7, Appendix B].

![Fig. 1: A block diagram of the scaled FHMM-based SCSS approach.](image-url)
5. EXPERIMENTS

5.1. Experimental Setup

Speech files considered for the experiments are selected from the database presented in [9]. Twelve speakers are selected to form the mixtures of female-female, male-male, and female-male pairs. After mixing speech files, ten female-female, ten male-male, and ten female-male mixtures (observations) are obtained. The speech files are mixed at the TIRs of 0, 3, 6, 9, 12, and 15 dB such that 180 different observations are generated for the experiments. Throughout the experiments, a Hamming window of length 32 msec with the frame shift equal to 10 msec is used to segment the speech files. The proposed technique is compared with a similar approach in which vector quantization (VQ) [7] is used for modeling instead of FHMM.

For each speaker, 100 sentences are used for FHMM and VQ training. The windowed training speech files are transformed into the log frequency domain using a 256-point discrete Fourier transform (\( D = 256 \)), resulting in log spectral vectors of dimension 129. For VQ modeling, the LBG VQ algorithm with binary splitting initialization is used to construct a 64-entry codebook (\( K = 64 \)) for each speaker. For FHMM modeling, we used the Baum-Welch method to estimate the FHMM parameters. The number of states is set to 64 (\( K = 64 \)). The initial estimates of the FHMM parameters are obtained from the VQ training. The Baum-Welch method iterates until either the difference between the current and previous log likelihood is equal or less 0.00001 or the maximum number of iterations is reached (we set the maximum number of iterations to 15).

5.2. Results

In order to evaluate the separation performance of the proposed techniques, we use the signal-to-noise ratio (SNR) between the estimated and original speech files. SNR results are shown in Fig. 2-Fig. 7. In Fig. 2, Fig. 4, and Fig. 6, we show the SNR versus \( \theta \) averaged over 10 separated target speech files for female-female, male-male, and female-male mixtures, respectively. For: scaled-FHMM (\( \bigcirc \) line), scaled-FHMM (with actual \( \theta \) (\( \bigtriangleup \) line), scaled-VQ (\( \triangleleft \) line), scaled-VQ (\( \odot \) line), scaled-VQ (\( \Theta \) line), HMM (\( \bigtriangledown \) line), and VQ (\( \Delta \) line). Also, Fig. 3, Fig. 5, and Fig. 7 show the same respective results, but for the separated interference signals. From the figures, several comparisons can be made. These observed trends hold for all three types of mixtures. The first observation is that as \( \theta \) increases the SNRs for the target signals increase as well and, on the contrary, the SNRs decrease for the interference signals. This behavior of the SNR curves versus \( \theta \) is quite expected since a signal with higher power can be better separated from the weaker one. Comparing the scaled FHMM technique with actual and estimated \( \theta \) (\( \bigcirc \) lines with \( \Box \) lines), we see that the SNR results are almost the same. In fact, even slight improvements are seen for the scaled FHMM with \( \theta \) estimated using quadratic optimization. The same trends also hold for the scaled VQ with actual and estimated \( \theta \) (\( \triangleleft \) lines with (\( \odot \) lines). It should be noted that an ideal \( \theta \) applied in the separation process is not be necessarily the actual \( \theta \). The actual \( \theta \) is the best choice if the original, rather than the model-supplied, log spectral vectors are used for separation.

Comparing SNR results obtained from the scaled-FHMM with scaled-VQ technique (\( \bigcirc \) lines with (\( \triangleleft \) lines), we observe that the scaled-FHMM technique outperforms the scaled-VQ for both target and interference signals in all three type of mixtures.

We also compare SNR results obtained from the scaled-FHMM with HMM and scaled-VQ with VQ (\( \bigcirc \) and \( \triangleleft \) lines with (\( \bigtriangledown \) and (\( \Delta \) lines). The results show that the scaled versions of HMM and VQ techniques significantly outperform the non-scaled ones. The improvement is quite palpable for \( \theta s > 6 \text{ dB} \). The results confirm that the model-based SCSS techniques fail to separate the speech signals when the test samples have energies substantially different from those used in the training data set. For the compared techniques, separation of interference signals at \( \theta s > 6 \text{ dB} \) is a difficult task as the separated interference signal has very poor quality, showing that solving this problem remains a challenge for future studies.

![Fig. 2: SNR versus \( \theta \) averaged over 10 separated target speech files from female-female mixtures.](image)

![Fig. 3: SNR versus \( \theta \) averaged over 10 separated interference speech files from female-female mixtures.](image)

6. CONCLUSIONS

Ignoring the difference between speakers’ gain significantly degrades the performance of model-based single channel speech techniques. We proposed a new technique to compensate the gain difference in factorial HMMs based separation which has been recognized the most accurate and robust modeling approach currently available. The experimental results showed that this new model, scaled FHMM, significantly outperforms not only unscaled ap-
Fig. 4: SNR versus $\theta$ averaged over 10 separated target speech files from male-male mixtures.

Fig. 5: SNR versus $\theta$ averaged over 10 separated interference speech files from male-male mixtures.

Fig. 6: SNR versus $\theta$ averaged over 10 separated target speech files from female-male mixtures.

Fig. 7: SNR versus $\theta$ averaged over 10 separated interference speech files from female-male mixtures.

Figures 4, 5, 6, and 7 show the Signal-to-Noise Ratio (SNR) versus $\theta$ for different conditions. The graphs compare the performance of various models (Scaled-HMM, Scaled-HMM (actual $T$), Scaled-VQ, Scaled-VQ (actual $T$), HMM, VQ) under different conditions: male-male target and interference, and female-male target and interference.

7. REFERENCES


