Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space

Tuomas Virtanen

Tampere University of Technology
Institute of Signal Processing
tuomas.virtanen@tut.fi

Abstract

This paper proposes an algorithm for the recognition and separation of speech signals in non-stationary noise, such as another speaker. We present a method to combine hidden Markov models (HMMs) trained for the speech and noise into a factorial HMM to model the mixture signal. Robustness is obtained by separating the speech and noise signals in a feature domain, which discards unnecessary information. We use mel-cepstral coefficients (MFCCs) as features, and estimate the distribution of mixture MFCCs from the distributions of the target speech and noise. A decoding algorithm is proposed for finding the state transition paths and estimating gains for the speech and noise from a mixture signal. Simulations were carried out using speech material where two speakers were mixed at various levels, and even for high noise level (9 dB above the speech level), the method produced relatively good (60% word recognition accuracy) results. Audio demonstrations are available at www.cs.tut.fi/~tuomasv.

Index Terms: speech recognition, speech separation, factorial hidden Markov model.

1. Introduction

One of the major problems in automatic speech recognition (ASR) is the degraded performance when the target speech is interfered with a noise source. Methods which try to overcome this problem can be roughly divided into two categories: 1) those which try to estimate the clean speech waveform or its features from the noisy signal, i.e., which perform speech separation, and 2) those developed and trained to recognize noisy speech (see [1] for a review.) Non-stationary noise which has similar acoustic characteristics as the target speech is especially difficult for all methods. Since a good estimate of the interference is a requirement for its efficient suppression, an accurate acoustic modeling of the noise can provide an increase in the recognition quality.

The ASR system proposed in this paper achieves robustness to interference by doing the separation in the feature domain. The speech and noise signals are modeled with hidden Markov models (HMMs), which are estimated beforehand from material where the signals are present in isolation. We do not commit ourselves to a specific method for training the HMMs, but Section 5 presents one possibility for this. To model the noisy speech signal, the source-specific HMMs are combined into a factorial HMM, as explained in Section 2. The likelihoods for noisy observation vectors are calculated from speech and noise distributions, as explained in Section 2.1. The recognition is done by finding the most likely state transition paths jointly for the speech and noise using the algorithm proposed in Section 3. Section 4 presents a method for synthesizing the sources separately, and Section 5 presents the simulations.

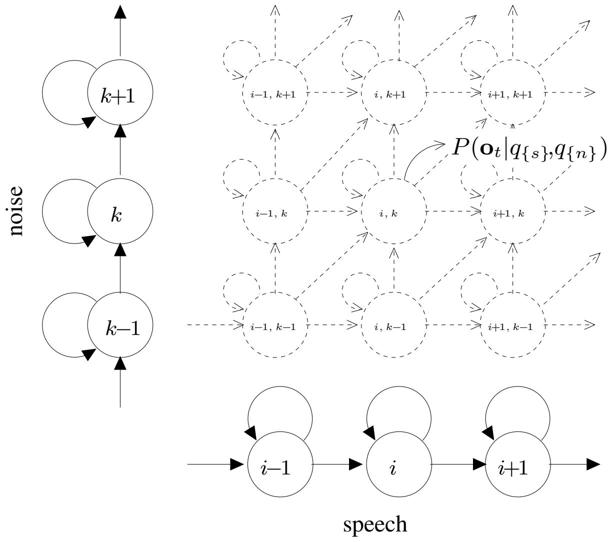

Figure 1: A factorial HMM consisting of Markov chains for the speech and noise (solid lines) can be synthesized into a HMM containing all the combinations of the speech and noise states (dotted lines).

2. Factorial Model for the Mixture Signal

When multiple sources are present simultaneously, the acoustic mixture signal is the superposition of the source signals. In this paper we consider only two sources, which are referred as "speech" and "noise", and labels $\{s\}$ and $\{n\}$ are used to refer to their parameters, respectively. Both signals are modeled with separate HMMs, which are assumed to be trained separately beforehand. The mixture signal is modeled with a factorial HMM [2] consisting of separate Markov chains for the speech and noise, which model the contribution of the speech and noise signals in the mixture, respectively.

The topology of the Markov chains is the same as the original HMMs, as illustrated in Figure 1. The transitions of the hidden state variables $q_{\{s\}}$ and $q_{\{n\}}$ of the speech and noise are statistically independent from each other, so that the state transition probability from state (i, k) to (j, l) is the product of the state transition probabilities of both chains:

$$P(q_{\{s\}}^{t+1} = j, q_{\{n\}}^{t+1} = l \mid q_{\{s\}}^{t} = i, q_{\{n\}}^{t} = k)$$
$$= P(q_{\{s\}}^{t+1} = j \mid q_{\{s\}}^{t} = i)\, P(q_{\{n\}}^{t+1} = l \mid q_{\{n\}}^{t} = k), \quad (1)$$

where t is the observation index.

The likelihood of an observation, however, depends on the state of the both chains, and it should be determined using the linear superposition of acoustic signals. In addition to factorial HMMs, also the term parallel mode combination has been used for this kind of approach by Gales and Young [3], who used single-state and two-state HMMs to model the noise. In our application the left-to-right topology is most suitable since the noise is assumed to be structured, but the topology can be fully connected as well.

Several algorithms based on factorial HMMs simplify the mixing process by modeling the power spectrum of the mixture signal at a certain time-frequency point as a maximum of the source power spectra, plus a noise term [4]. The approximation leads to a closed-form expression for the distribution of the sum of speech and noise [5, 6] when the power spectra are used as observations. In ASR, however, the power spectrum representation lacks invariancy to some features such as pitch, and therefore the approximation requires a large amount of components to model different phoneme-pitch combinations.

2.1. Likelihoods for MFCCs

Mel-frequency cepstral coefficients (MFCCs) are commonly used to parameterize the rough shape of the spectrum in ASR. For a single frame of speech they are calculated by measuring the power within mel-frequency bands, taking the logarithm, and decorrelating the resulting vector by the discrete cosine transform (DCT). The above-explained max-approximation of the sum is not reasonable for MFCCs, and therefore estimating the likelihood of a mixture observation requires more accurate modeling of the mixing process when MFCCs are used as features.

The probability density function of an observed MFCC vector $\mathbf{o}_t$ is commonly modeled by a Gaussian mixture model (GMM), the parameters of which are different for each state. In the following we derive the distribution for the mixture signal MFCCs of the factorial HMM state (i, k) using the GMMs of the target state i and noise state k. The process has to be repeated for all i and k to get distributions of all state combinations.

For simplicity, we formulate the distribution here for a single-component GMM, which is the normal distribution $\mathcal{N}$. For clean speech, the likelihood of an observation $\mathbf{o}_t$ is given as

$$P(\mathbf{o}_t) = \mathcal{N}(\mathbf{o}_t \mid \boldsymbol{\mu}_{\{s\}}^{\{m\}}, \boldsymbol{\Sigma}_{\{s\}}^{\{m\}}), \quad (2)$$

where $\boldsymbol{\mu}_{\{s\}}^{\{m\}}$ is a mean vector and $\boldsymbol{\Sigma}_{\{s\}}^{\{m\}}$ is a covariance matrix of the distribution, and label $\{m\}$ is used to denote MFCC-domain parameters. The formulation can be extended for multi-component GMMs, as explained later.

To enable estimating the likelihood of a mixture observation from the target and noise distributions, the speech and noise MFCC distributions have to be transformed to the power spectral domain, where the distribution of their superposition can be estimated. From the power spectral domain the parameters are again transformed back to the MFCC domain, as illustrated in Figure 2. The power-spectral domain mean and covariance of the speech are calculated using the steps below. The method is essentially the same as presented in [3], extended for multiple states, and the proposed decoding algorithm is original.

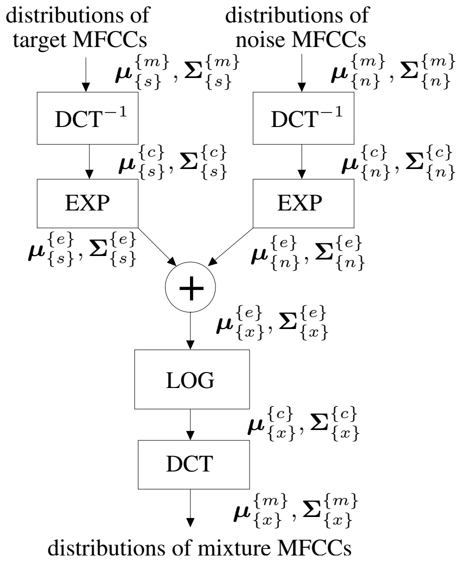
distributions of target MFCCs
distributions of noise MFCCs
$\boldsymbol{\mu}_{\{s\}}^{\{m\}}, \boldsymbol{\Sigma}_{\{s\}}^{\{m\}}$
$\boldsymbol{\mu}_{\{n\}}^{\{m\}}, \boldsymbol{\Sigma}_{\{n\}}^{\{m\}}$
DCT^{-1}
DCT^{-1}
$\boldsymbol{\mu}_{\{s\}}^{\{c\}}, \boldsymbol{\Sigma}_{\{s\}}^{\{c\}}$
$\boldsymbol{\mu}_{\{n\}}^{\{c\}}, \boldsymbol{\Sigma}_{\{n\}}^{\{c\}}$
EXP
EXP
$\boldsymbol{\mu}_{\{s\}}^{\{e\}}, \boldsymbol{\Sigma}_{\{s\}}^{\{e\}}$
$\boldsymbol{\mu}_{\{n\}}^{\{e\}}, \boldsymbol{\Sigma}_{\{n\}}^{\{e\}}$
+
$\boldsymbol{\mu}_{\{x\}}^{\{e\}}, \boldsymbol{\Sigma}_{\{x\}}^{\{e\}}$
LOG
$\boldsymbol{\mu}_{\{x\}}^{\{c\}}, \boldsymbol{\Sigma}_{\{x\}}^{\{c\}}$
DCT
$\boldsymbol{\mu}_{\{x\}}^{\{m\}}, \boldsymbol{\Sigma}_{\{x\}}^{\{m\}}$
distributions of mixture MFCCs

Figure 2: Calculation of the mixture MFCC distributions for a single state.

First, the logarithm of the powers in mel-frequency bands is obtained by taking an inverse DCT of the MFCCs, which can be formulated by multiplication by inverse of the DCT matrix $\mathbf{C}$. Therefore, the distribution of the logarithms of the powers is also a normal distribution, with mean

$$\boldsymbol{\mu}_{\{s\}}^{\{c\}} = \mathbf{C}^{-1}\boldsymbol{\mu}_{\{s\}}^{\{m\}} \quad (3)$$

and covariance

$$\boldsymbol{\Sigma}_{\{s\}}^{\{c\}} = \mathbf{C}^{-1}\boldsymbol{\Sigma}_{\{s\}}^{\{m\}}(\mathbf{C}^{-1})^{T}, \quad (4)$$

where label $\{c\}$ denotes cepstral-domain parameters.

The energies within mel-frequency bands are obtained by taking the exponential function of the log-energies. Their mean $\boldsymbol{\mu}_{\{s\}}^{\{e\}}$ and covariance $\boldsymbol{\Sigma}_{\{s\}}^{\{e\}}$ can be shown to be [3]

$$\mu_{i\{s\}}^{\{e\}} = \exp(\mu_{i\{s\}}^{\{c\}} + \Sigma_{ii\{s\}}^{\{c\}}/2), \quad (5)$$

and

$$\Sigma_{ij\{s\}}^{\{e\}} = \mu_{i\{s\}}^{\{c\}}\mu_{j\{s\}}^{\{c\}}[\exp(\Sigma_{ij\{s\}}^{\{c\}}) - 1], \quad (6)$$

where label $\{e\}$ denotes the power-spectrum domain parameters, μ_i denotes the i^{th} element of vector $\boldsymbol{\mu}$, and Σ_{ij} denotes the $(i, j)^{\text{th}}$ element of matrix $\boldsymbol{\Sigma}$.

Equations from (3) to (6) are applied to both the speech and noise, to result in their power-spectral means $\boldsymbol{\mu}_{\{s\}}^{\{e\}}$ and $\boldsymbol{\mu}_{\{n\}}^{\{e\}}$ and covariances $\boldsymbol{\Sigma}_{\{s\}}^{\{e\}}$ and $\boldsymbol{\Sigma}_{\{n\}}^{\{e\}}$, respectively. The expectation value for the energy of the superposition of statistically independent signals is the sum of their energies. Therefore, the mean vector of the mel-energies of the mixture is

$$\boldsymbol{\mu}_{\{x\}}^{\{e\}} = g_{\{s\}}\boldsymbol{\mu}_{\{s\}}^{\{e\}} + g_{\{n\}}\boldsymbol{\mu}_{\{n\}}^{\{e\}}, \quad (7)$$

where the label $\{x\}$ denotes mixture parameters, and $g_{\{s\}}$ and $g_{\{n\}}$ are the gains of the speech and noise, respectively. They are used to accommodate possible level differences between the training and target material, and their estimation is explained in Section 3. Similarly, the covariances of statistically independent signals sum linearly, so that that covariance matrix of the mel-energies of the mixture signal is

$$\mathbf{\Sigma}^{\{e\}}_{\{x\}} = g^2_{\{s\}}\mathbf{\Sigma}^{\{e\}}_{\{s\}} + g^2_{\{n\}}\mathbf{\Sigma}^{\{e\}}_{\{n\}}. \quad (8)$$

Since the distribution of the logarithm of mel-energies is normal, the distribution of mel-energies is log-normal. There is no closed-form solution for the distribution of the sum of log-normally distributed random variables [7]. For a given observation sequence the probabilities could be estimated, for example, by numerical integration, but this is not feasible because of the computational complexity. However, it has been observed that the sum can be rather well be approximated by another log-normal distribution [7] and there are several methods for calculating its parameters.

We use the method proposed by Gales and Young [3], which is summarized as follows: first, the mean and covariance of the log-energies of the mixture are given as

$$\mu^{\{c\}}_{i\{x\}} = \log\left(\mu^{\{e\}}_{i\{x\}}\right) - \frac{1}{2}\log\left(\frac{\Sigma^{\{e\}}_{ii\{x\}}}{\left(\mu^{\{e\}}_{i\{x\}}\right)^2} + 1\right) \quad (9)$$

and

$$\Sigma^{\{c\}}_{ij\{x\}} = \log\left(\frac{\Sigma^{\{e\}}_{ij\{x\}}}{\mu^{\{e\}}_{i\{x\}}\mu^{\{e\}}_{j\{x\}}} + 1\right). \quad (10)$$

The mean and variance of the MFCCs of the sum are obtained by taking the DCT, so that:

$$\boldsymbol{\mu}^{\{m\}}_{\{x\}} = \mathbf{C}\boldsymbol{\mu}^{\{c\}}_{\{x\}} \quad (11)$$

and

$$\mathbf{\Sigma}^{\{m\}}_{\{x\}} = \mathbf{C}\mathbf{\Sigma}^{\{c\}}_{\{x\}}\mathbf{C}^T. \quad (12)$$

Finally, the likelihood that state (i, k) produces observation $\mathbf{o}_t$ is the value of the normal distribution $\mathcal{N}(\mathbf{o}_t \mid \boldsymbol{\mu}^{\{m\}}_{\{x\}}, \mathbf{\Sigma}^{\{m\}}_{\{x\}})$, where $\boldsymbol{\mu}^{\{m\}}_{\{x\}}$ and $\mathbf{\Sigma}^{\{m\}}_{\{x\}}$ are calculated using Equations (3) to (12). To reduce the computational complexity and increase numerical robustness, we used only diagonal of the resulting covariance matrix. In general this does not have significant effect on the results [3].

The method can be extended to n-component GMMs by calculating the distribution for the mixture MFCCs individually for each GMM-component pair of the target and noise, which are then summed to yield an n^2-component GMM of the mixture signal. Similar formulas can also be derived for delta-MFCCs, but they are not presented because of space limitation.

3. Decoding Algorithm

The objective of the decoding algorithm is to find state transition paths of the target and noise so that the total likelihood, which is the sum of the observation likelihoods and state transition probabilities, is maximized. We assume that the HMMs of the target and noise are not necessarily trained with material where the signal levels are equal to those in the mixture signal, so that the gains $g_{\{s\}}$ and $g_{\{n\}}$ also have to be estimated. Because of the complexity of the model, finding the global optimum is not possible, and therefore we propose a greedy algorithm consisting of the following steps:

1. Find the best state transition path and $g_{\{s\}}$ for the speech alone. This is done by a one-dimensional optimization [8] $g_{\{s\}}$, where the Viterbi algorithm is used to estimate the optimal state transition path for each tested value of $g_{\{s\}}$.
2. Initialize $g_{\{n\}} = \alpha g_{\{s\}}$, where the fixed scalar α was chosen to have value 0.3.
3. Find the best state transition paths of the noise and speech simultaneously. The estimation algorithm can be viewed as an extension of the Viterbi algorithm, where the Markov chains of speech and noise are synthesized into a single HMM, as done by Gales and Young [3]. To reduce the computational complexity, the speech chain is allowed to use only those states which were used in its previously most likely state transition path.
4. While keeping the state transition paths fixed, optimize $g_{\{s\}}$ and $g_{\{n\}}$ using the Nelder-Mead [9] algorithm.
5. Find the best state transition paths for the noise and speech simultaneously. To reduce the computational complexity, the noise chain is allowed to use only those states which were used in the previously most likely path.
6. While keeping the state transition paths fixed, optimize $g_{\{s\}}$ and $g_{\{n\}}$ using the Nelder-Mead algorithm.

The steps 3-6 are repeated until the likelihood of the model does not increase. For e.g. 150 observations and 8500 states for different HMMs in both the speech and noise chains the algorithm takes a couple of iterations to converge, which takes several minutes on a 3.2 GHz PC when implemented in Matlab.

4. Synthesis

To allow post-processing and quality evaluation by listening, a method for synthesizing the speech and noise signals separately was also developed. The synthesis is done by filtering the mixture signal by a time-varying Wiener filter, which is designed using the energies predicted by the target and noise chains.

For each frame the filter is designed as follows. Let use denote the state of the most likely state transition path in frame t by indices (i, k). First, the mean and covariance of the GMMs of speech state i and noise state k of the original GMMs are used to calculate the mean energy vectors $\boldsymbol{\mu}^{\{e\}}_{\{s\}}$ and $\boldsymbol{\mu}^{\{e\}}_{\{n\}}$ using Equations (3) to (5).

The power response W_i of the Wiener filter for the speech at mel-frequency i is given as

$$W^{\{s\}}_i = \frac{g_{\{s\}}\mu^{\{e\}}_{i\{s\}}}{g_{\{s\}}\mu^{\{e\}}_{i\{s\}} + g_{\{n\}}\mu^{\{e\}}_{i\{n\}}} \quad (13)$$

and the Wiener filter for the noise as $1 - W^{\{s\}}_i$. The filtering can be implemented by taking the discrete Fourier transform (DFT) of the frame t, multiplying each bin of the resulting spectrum by the square root of the power response of the Wiener filter at corresponding mel-band i, taking the inverse DFT, and combining adjacent frames using overlap-add. The method produces speech signals where the noise is significantly suppressed, and no significant artefacts are introduced on the speech. Audio demonstrations are available at `www.cs.tut.fi/~tuomasv`.

5. Simulations

The system was tested using the material of the speech separation challenge.[1] The acoustic material was drawn from the GRID corpus [10] consisting of six-word sentences where the total number of different words is 52, spoken by 34 different speakers. In each test signal, two speakers were mixed at relative levels ranging from -9 dB to 6 dB, the number of signals per each mixing level being 600. The target of the challenge is to recognize a letter and a digit spoken by the speaker saying the word “white”. The identities or the relative levels of the speakers were not used in the recognition.

5.1. Training

Speaker-specific HMMs were trained using similar sentences spoken in isolation. Annotations where the word-level transcription and the acoustic signal were aligned [10] were used to segment the signals into words. Twenty-four MFCCs were calculated within 30 ms windows with 50% overlap, and a HMM per each word per each speaker was trained using the Baum-Welch algorithm. The number of states per word was two times the number of phonemes, resulting in between 4 and 10 states per word. To minimize the computational complexity, we used single-component GMMs to model the MFCC distributions.

Sentence-level HMMs were built by concatenating the word HMMs. Since the speaker identities were not known, the final HMM for the speech was obtained by putting the sentence-level HMMs of each speaker in parallel. The noise signals in the simulations were drawn from the same database as the speech, and therefore the noise HMM was exactly the same as the speech HMM.

5.2. Recognition

The state transition paths for the speech and noise were determined using the algorithm described in Section 3, and the words spoken by both speakers were inferred from the paths. The speaker who was recognized to say the word “white” was regarded as the target speaker. If none or both speakers said “white”, the speaker with larger gain was regarded as the target.

The word recognition accuracy was measured as the ratio of the number of correctly recognized words per the total number of words to be recognized. These were calculated separately for each mixing level, and also separately for cases where the genders of the talkers were the same or different. The challenge includes also test cases where the speech and noise originate from the same speaker, and the averages were calculated separately also for these cases.

5.3. Results

The average word recognition rates are shown in Table 1. The method is able to produce relatively good performance even for high noise levels, the average quality decreasing gradually as the noise level increases. Mixtures where the speaker and noise identities were different were more easy to recognize. When the target and noise speaker identities were the same, increasing the noise level increased the average recognition rate, since the level difference aided to distinguish between the target and noise.

In these simulations the noise HMMs were trained using isolated noise signals; estimating the noise HMM for real-world mixtures may not be as straightforward. When the speech HMM alone was used to recognize clean speech signals, we obtained recognition accuracy of 97%. The proposed factorial HMM method was not allowed to assign zero gains for the noise, decreasing its performance in the clean condition.

Table 1: Average word recognition rates (%) for different speech-to-noise ratios (SNR), and without noise (clean).

SNR	Same speaker	Same gender	Diff. gender	Avg.
clean	88	87	89	88
6 dB	63	86	82	76
3 dB	57	83	82	73
0 dB	44	76	78	65
-3 dB	36	74	73	60
-6 dB	45	71	71	61
-9 dB	46	70	65	60

6. Conclusions

The proposed factorial HMM produces applicable results in the recognition and separation of speech in the interference of another speaker. The simulations show that separation of simultaneous speech signals is possible using only the rough shape of spectrum parameterized by MFCCs and temporal structure modeled by HMMs. The applied method for calculating the MFCCs of the mixture signal is a good alternative for the widely-used max-approximation in factorial HMMs.

7. References

[1] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, 2005.

[2] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Machine Learning*, vol. 29, 1997.

[3] M. J. F. Gales and S. J. Young, “Cepstral parameter compensation for HMM recognition in noise,” *Speech Communication*, vol. 12, 1993.

[4] S. T. Roweis, “Factorial models and refiltering for speech separation and denoising,” in *EuroSpeech*, Geneva, Switzerland, 2003.

[5] A. Nadas and M. A. Picheny, “Speech recognition using noise-adaptive prototypes,” *IEEE Transactions on Speech and Audio Processing*, vol. 37, no. 10, 1989.

[6] A. P. Varga and R. K. Moore, “Hidden Markov model decomposition of speech and noise,” in *IEEE Int. Conf. on Audio, Speech and Signal Processing*, Albuquerque, USA, 1990.

[7] N. C. Beaulieu, A. A. Abu-Dayya, and P. J. McLane, “Estimating the distribution of a sum of independent lognormal random variables,” *IEEE Trans. on Comm.*, vol. 43, 1995.

[8] R. P. Brent, *Algorithms for Minimization without Derivatives*. Mineola, New York: Dover Publications, 1973.

[9] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, “Convergence properties of the Nelder-Mead simplex method in low dimensions,” *SIAM J. Optim.*, vol. 9, 1998.

[10] M. P. Cooke, J. Barker, S. P. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *J. Acoust. Soc. Am.*, 2005, submitted.

[1] www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm