# Single Channel Speech Separation Using Layered Hidden Markov Models

**Anonymous Author(s)**
Affiliation
Address
City, State/Province, Postal Code, Country
email

## Abstract

We present a model-based system capable of separating and recognizing speech of two speakers from a single-channel recording. The system uses models of speech that capture dynamics at two levels: an acoustic level that models the continuity of the spectrum, and a grammatical level that models linguistic constraints from the phone level up to phrases. We present speech recognition experiments demonstrating that the full system performs close to humans overall. Remarkably the system exceeds human recognition performance in 0dB conditions. Since the pattern of performance across conditions is quite different for humans, we hypothesize that the auditory system uses different strategies than our model.

## 1 Introduction

Listening to and understanding the speech of two people when they talk simultaneously is a difficult task and has been considered one of the most challenging problems for automatic speech recognition, especially when only a single-channel signal is available.

To address this problem, single-channel speech separation has been attempted using Gaussian mixture models (GMMs) on individual frames of acoustic features. However such models tend to perform well only when speakers are of different gender or have rather different voices [1]. When speakers have similar voices, speaker-dependent mixture models cannot unambiguously identify the component speakers. In such cases it is helpful to model the temporal dynamics of the speech. Several models in the literature have attempted to do so either for recognition [2, 3] or enhancement [4, 5] of speech. Such models have typically been based on a discrete state hidden Markov model (HMM) operating on a frame-based acoustic feature vector.

One of the challenges of such modeling is that speech contains patterns at different levels of detail, that evolve at different time-scales. For instance, two major components of the voice are the excitation, which consists of pitch and voicing, and the filter, which consists of the formant structure due to the mouth position. The pitch appears in the short-time spectrum as a closely-spaced harmonic series of peaks, whereas the formant structure has a smooth frequency envelope. The formant structure and voicing are closely related to the phoneme being spoken, whereas the pitch evolves somewhat independently of the phonemes during voiced segments.

At small time-scales these processes evolve in a somewhat predictable fashion, with relatively smooth pitch and formant trajectories, interspersed with sharper transients. If we begin with a Gaussian mixture model of the log spectrum, we can hope to capture something about the dynamics of speech by just looking at pair-wise relationships between the acoustic states ascribed to individual frames of speech data.

In addition to these low-level acoustical constraints, there are linguistic constraints that describe the dynamics of syllables, words, and sentences. These constraints depend on context over a longer time-scale and hence cannot be modeled by pair-wise relationships between acoustic states. In speech recognition systems such long-term relationships are handled using concatenated left-to-right models of context-dependent phonemes, that are derived from a grammar or language model.

Typically models in the literature have focused on only one type of dynamics. Here we explore the combination of both the low-level acoustic dynamics with the high-level grammatical constraints. We compare three levels of dynamic constraints: simple GMM models of the log spectrum with no dynamics, acoustic-level HMM dynamics, and a layered combination of acoustic-level and grammar-level dynamics. The models are combined at the observation level using a nonlinear model known as Algonquin, which models the sum of log-normal spectrum models. Inference on the state level is carried out using an iterative two-dimensional Viterbi decoding scheme.

The system is composed of the three components: a speaker identification and gain estimation component, a signal separation component and a speech recognition system. In this paper we focus on the signal separation component, which is composed of the acoustic and grammatical models. The details of the speaker ID and gain estimation along with the speech recognition system are provided in a system-level paper *[obfuscated for review]*.

The task we addressed is provided by the PASCAL[1] Speech Separation Challenge (SSC) [6], which provides standard training, development, and test datasets of single-channel speech mixtures following an arbitrary but simple grammar. In addition, the challenge organizers have conducted human-listening experiments to provide an interesting baseline for comparison of computational techniques.

Using both acoustic and grammar-level dynamics on this task, our system produces astonishing results. The system is often able to extract two utterances even when they are both recordings of the same speaker. In addition in conditions near 0dB signal-to-noise ratio (SNR), the system significantly exceeds human listener performance according to experiments conducted by the challenge organizers. Overall the system significantly improves on other methods used in the same challenge.

By the end of the year we will be able to compare our results with those of the other participants in the challenge. In addition we hope to address several questions raised by the experiments completed so far. First it is important to understand which constraints are more important, the grammar versus the acoustic, although preliminary results show that both are necessary for the best results. Second, we are exploring alternate methods of learning and inference in the model. Since the model has an intractable loopy structure, two alternatives are loopy belief propagation, and structured variational approximation. Thirdly, it is important to know how well such a model can perform if the test speakers do not occur in training data. Experiments underway will address these issues in the near future.

## 2 Source Models and Likelihood Estimation

The Speech Separation Challenge involves recognizing speech in files that are mixtures of two component signals. Each of the component signals, $x_t^a$ and $x_t^b$ for speaker a and b are modeled by a conventional continuous observation hidden Markov model (HMM) with Gaussian mixture models (GMM) for representing the observations. The main difference between our model and that of a standard recognizer is that observations are in the log-power spectrum domain. Hence, given an HMM state $s^a$ of speaker $a$, the distribution for the log spectrum vector $\mathbf{x}^a$ is modeled as $p(\mathbf{x}^a|s^a) = N(\mathbf{x}^a; \mu_{s^a}, \Sigma_{s^a})$.

The model for mixed speech in the time domain is (omitting the channel) $y_t = x_t^a + x_t^b$ where $y_t$ denotes the mixed signal at time $t$. We approximate this nonlinear relationship in the log spectrum domain as

$$p(\mathbf{y}|\mathbf{x}^a, \mathbf{x}^b) = N(\mathbf{y}; \ln(\exp(\mathbf{x}^a) + \exp(\mathbf{x}^b)), \mathbf{\Psi}) \quad (1)$$

where $\mathbf{\Psi}$ is introduced to model the error due to the omission of phase, and time has been omitted for simplicity. The combination of the two models can be depicted in a graph as in Figure 1.

[1]PASCAL (Pattern Analysis, Statistical modelling and ComputAtional Learning) is a project of the European Commission's Information Society Technologies, and hosts a series of open challenges to the research community. See `http://www.pascal-network.org/Challenges/`.


(a) Combining GMMs (b) Simplified Model

Figure 1: Graph of GMM model combination for two sources. In (a), the full GMM models are shown. In (b), the $x^a$ and $x^b$ have been integrated out.

The joint distribution of the two sources, their state and the observation is

$$p(\mathbf{y}, \mathbf{x}^a, \mathbf{x}^b, s^a, s^b) = p(\mathbf{y}|\mathbf{x}^a, \mathbf{x}^b)p(\mathbf{x}^a|s^a)p(\mathbf{x}^b|s^b)p(s^a)p(s^b). \quad (2)$$

### 2.1 Likelihood Estimation

Unlike a traditional recognizer, we must take into account the joint evolution of the two signals simultaneously. Hence we need to evaluate the joint observation likelihood $p(\mathbf{y}|s^a, s^b)$ at every time step.

The Newton-Laplace method can be used to approximate the joint posterior computed from Eqn. (2) with a weighted Gaussian. Once the joint distribution has been approximated, $p(\mathbf{y}|s^a, s^b)$ can be found by integrating out $x^a$ and $x^b$. This method of acoustic model combination is known as Algonquin [1]. We can find the minimum mean squared error (MMSE) estimate for $\mathbf{x}^a$ and $\mathbf{x}^b$ by taking the expected value $E[x^a, x^b|y]$, where the expected value is with respect to (2). Likewise the maximum a posteriori (MAP) estimates for $\mathbf{x}^a$ and $\mathbf{x}^b$ are given by $E[x^a, x^b|y, \hat{s^a}, \hat{s^b}]$, where $(\hat{s^a}, \hat{s^b}) = \arg\max_{s^a, s^b} p(s^a, s^b|y)$.

We used 256 component Gaussian mixture models (GMM) to model the acoustic space of each speaker. In this case, the evaluation of $p(\mathbf{y}|s^a, s^b)$ requires the evaluation of $256^2$ or over 65k state combinations.

### 2.2 Fast Likelihood Estimation

In order to speed up the evaluation of the joint observation likelihood, we employed both *band quantization* of the component GMMs and joint-state pruning. This gave several orders of magnitude speedup over the brute force approach.

Band quantization involves approximating each of the $D$ Gaussians of each model with a shared set of $d$ Gaussians, where $d \ll D$, in each of the 319 frequency bands. It relies on the use of a diagonal covariance matrix, so that $p(x^a|s^a) = \prod_f N(x^a_f; \mu_{f,s^a}, \sigma^2_{f,s^a})$, where $\sigma^2_{f,s^a}$ are the diagonal elements of covariance matrix $\Sigma_{s^a}$. The mapping $M_f(s_i)$ associates each of the $D$ Gaussians with one of the $d$ Gaussians in frequency band $f$. Now $\hat{p}(x^a|s^a) = \prod_f N(x^a_f; \mu_{f,M_f(s^a)}, \sigma^2_{f,M_f(s^a)})$ is used as a surrogate for $p(x^a|s^a)$. Under this model the $d$ Gaussians are chosen to minimize the KL-distance $D(p(x^a|s^a)||\hat{p}(x^a|s^a))$, and likewise for $s^b$. Then in each frequency band, only $d \times d$, instead of $D \times D$ combinations of Gaussians have to be evaluated to compute $p(\mathbf{y}|s^a, s^b)$. In our case, $d = 8$ and $D = 256$, so this saves over three orders of magnitude of computation time.

Only a handful of $s^a, s^b$ combinations are required to adequately explain the observation. By pruning the total number of combinations down to a smaller number we can speed up MMSE estimation of the components signals as well as the temporal inference. In the experiments reported here, we pruned down to 256 combinations.

The *max* approximation [5] provides an efficient if less accurate approximation to the joint observation likelihood. The max approximation assumes $p(\mathbf{y}|s^a, s^b) = p_{x^a}(\mathbf{y}|s^a)$ if the mean $\mu^a$ of $x^a$ is larger than the mean $\mu^b$ of $x^b$ and $p(\mathbf{y}|s^a, s^b) = p_{x^b}(\mathbf{y}|s^b)$ otherwise.

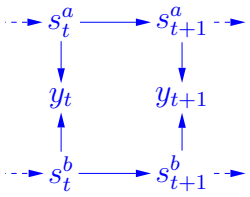
We relied on the max approximation for speaker identification and gain estimation and band-quantization followed by the Algonquin method on the pruned states for signal separation. The effect of these speedup methods on accuracy will be reported in a future publication.

## 3 Models of Temporal Dynamics

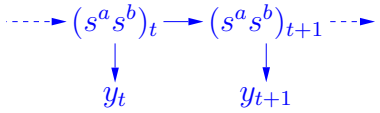
In a traditional speech recognition system, speech dynamics are captured by state transition probabilities. We took this approach and incorporated both *acoustic dynamics* and *grammar dynamics* via state transition probabilities.

### 3.1 Acoustic dynamics

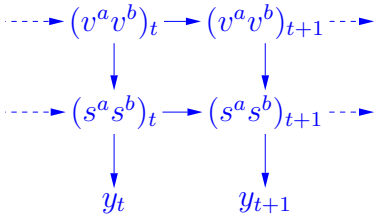
To capture acoustic level dynamics, which directly models the dynamics of the log-spectrum, we estimated transition probabilities between the states of the 256 component GMM models for each speaker. The acoustic dynamics of the two independent speakers are modeled by state transitions $p(s^a_{t+1}|s^a_t)$ and $p(s^b_{t+1}|s^b_t)$ for speaker a and b respectively as shown in Figures 2(a) and 2(b). Hence, for each speaker $c$, we estimated a $256 \times 256$ component transition matrix $A_c$.




(a) Acoustic HMM Combination




(b) Cartesian Product Acoustic HMM




(c) Layered Grammar + Acoustic Model

Figure 2: Graph of acoustic HMM model for two sources. In (a), the two state chains are shown separately. In (b), the $s^a$ and $s^b$ are combined into a Cartesian product state $(s^a s^b)$. In (c) a Cartesian product of two grammars $v^a$ and $v^b$ has been added on top of the acoustic state sequence. Note that this makes the graphical model loopy.

### 3.2 Grammar dynamics

The grammar dynamics are modeled by grammar state transitions, $p(v^c_{t+1}|v^c_t)$, which consist of left-to-right phone models. The legal word sequences are given by the SSC grammar [6] and are modeled using pronunciations that map to three-state context-dependent phone models. The sequences of phone states for each pronunciation, along with self-transitions produce a Finite State

Graph (FSG). The state transitions derived from this graph are sparse in the sense that most state transition probabilities are zero.

For a given speaker, the FSG of our system has 506 grammar states $v$. We then model speaker dependent distributions $p(s^c|v^c)$ that associate the FSG states to the speaker dependent GMM model states. These are learned from training data where the grammar state sequences and GMM state sequences are known for each utterance. This two-layer combined model is depicted in Figure 2(c).

To combine the acoustic dynamics with the grammar dynamics, it was useful to avoid modeling the full combination of $s$ and $v$ states in the joint transitions $p(s^c_{t+1}|s^c_t, v_t)$. Instead we make a naive-Bayes assumption to approximate this as $\frac{1}{z}p(s^c_{t+1}|s^c_t)p(s^c_{t+1}|v_{t+1})$, where $z$ is the normalizing constant.

## 4 Inference

Inference in the system involves estimating the posterior probability of the joint state sequences of the target and filter models and then computing the MAP estimate of their log spectra. In the results reported here this inference is approximated by finding the most likely state sequences for target and masker. In our experiments we performed inference in three different conditions: *GMM* inference, *acoustic dynamics*, and *grammar dynamics*.

The GMM inference has no temporal dynamics and source estimates $E(\mathbf{x}^a|\mathbf{y})$ and $E(\mathbf{x}^b|\mathbf{y})$ are inferred using posteriors of Eqn. (2) and marginalizing over states $s^a$, $s^b$.

In the acoustic dynamics condition, the exact inference algorithm uses the 2D Viterbi search. The Viterbi algorithm estimates the maximum likelihood state sequence $s_{1..T}$ given the observations $x_{1..T}$. The complexity of the Viterbi search is $O(D^2 \cdot T)$ where $D$ is the number of states and $T$ is the number of frames. For producing MAP estimates of the 2 sources, we require a 2 dimensional Viterbi search which finds the most likely joint state sequences $s^a_{1..T}$ and $s^b_{1..T}$ given the mixed signal $y_{1..T}$ as was proposed in [2]. Surprisingly, this 2D Viterbi search is of complexity $O(D^3 \cdot T)$, and not $O(D^4 \cdot T)$. By exploiting the sparsity of the transition matrices and pruning the observation likelihoods, our implementation of 2D Viterbi search is faster than the underlying Algonquin likelihood computation.

Thus for the acoustic dynamics condition, we use with acoustic temporal constraints $p(s_t|s_{t-1})$ and likelihoods from Eqn. (2), to find the most likely joint state sequence $s_{1..T}$.

In the grammar dynamics condition we use the layered model of section 3.2. Exact inference is computationally complex because the full joint distribution of the grammar and acoustic states, $(v^a \times s^a) \times (v^b \times s^b)$ is required and is very large in number. Thus we perform approximate inference by alternating the 2D Viterbi search between two layers: the Cartesian product $s^a \times s^b$ of the acoustic state sequences and the Cartesian product $v^a \times v^b$ of the grammar state sequences. This is a useful factorization because the states $s^a$ and $s^b$ interact strongly with each other and similarly for $v^a$ and $v^b$. In fact, in the same-talker condition the corresponding states exhibit an exactly symmetrical distribution. The 2D Viterbi search breaks this symmetry on each factor. When evaluating each state sequence we hold the other chain constant, which decouples its dynamics and allows for efficient inference.

We are currently exploring the alternative approximate inference strategies for this model: loopy belief propagation and a structured variational EM algorithm. In addition, we are investigating learning the joint model rather than combining two separately trained layers.

Once the maximum likelihood joint state sequence is found we can infer the source log-power spectrum of each signal and reconstruct them as shown in [1]. Although in the case of the grammar dynamics inference in the system actually involves recognition of the words, we still found that a separately trained recognizer performed better, perhaps because the recognizer's acoustic features are better suited for recognition. Each of the two signals is thus decoded with a speech recognition system that incorporates Speaker Dependent Labeling (SDL) *[citation obfuscated for review]*. This method uses speaker dependent models for each of the 34 speakers. Instead of using the speaker identities provided by the speaker ID and gain module, we followed the approach for gender dependent labeling (GDL) described in [7]. This technique provides better results than if the true speaker ID is specified.

## 5 Speaker Identification and Gain Estimation

In the SSC task, the gains and identities of the two speakers were unknown at test time and were selected from a set of 34 speakers which were mixed at SNRs ranging from 6dB to -9dB. We used speaker-dependent acoustic models because of their advantages when separating different speakers. These models were trained on data with a narrow range of gains, so it is necessary for inference to match the models to the gains of the signals at test time. This means that we have to estimate both the speaker identities and the gain in order to successfully infer the source signals.

However, the number of speakers and range of SNRs in the test set makes it too expensive to consider every possible combination of models and gains. Instead, we developed an efficient model-based method for identifying the speakers and gains. The algorithm is based upon a very simple idea: identify and utilize frames that are dominated by a single source to determine what sources are present in the mixture. The output of this stage is a short list of candidates. The combination of candidates on the short-list that maximizes the probability of the mixture under a gain adaptive approximate EM procedure is then selected.

| | 6 dB | 3 dB | 0 dB | -3 dB | -6 dB | -9dB | All |
|---|---|---|---|---|---|---|---|
| ST | 100 | 100 | 100 | 100 | 100 | 99 | 99 |
| SG | 97 | 98 | 98 | 97 | 97 | 96 | 97 |
| DG | 99 | 99 | 98 | 98 | 97 | 96 | 98 |
| All | 99 | 99 | 99 | 98 | 98 | 97 | 98 |

Table 1: Speaker identification accuracy (percent) as a function of test condition and case on the SSC two-talker test set, for the presented source identification and gain estimation algorithm. ST-Same Talker, SG-Same Gender, DG-Different Gender.

Table 1 reports the speaker identification accuracy obtained on the SSC two-talker test set via this approach, when all combinations of the most probable source and the six most probable sources are considered (six combinations total), and the speaker combination maximizing the probabilility of the data selected. Over all mixture cases and conditions on the SSC two-talker test set we obtained greater than 98% speaker identification accuracy overall.

## 6 Experiments and Results

The SSC [6] utterances consists a simple command sentence taken at random from a simple grammar. An example utterance is *place white by R 4 now*. In each recording, one of the speakers says *white* while the other says *blue*, *red* or *green*. The task is to recognize the letter and the digit of the speaker that said *white*.

We decoded the two component signals under the assumption that one signal contains white and the other does not, and vice versa. We then used the association that yielded the highest combined likelihood.

Log-power spectrum features were computed at a 15 ms rate. Each frame was of length 40 ms and a 640 point FFT was used, and the DC component was discarded, producing a 319-dimensional log-power-spectrum feature vector.

| | 6 dB | 3 dB | 0 dB | -3 dB | -6 dB | -9dB | All |
|---|---|---|---|---|---|---|---|
| ST | 29 | 42 | 47 | **47** | 46 | 55 | 44.3 |
| SG | 8 | **10** | **13** | **13** | **15** | 30 | **14.8** |
| DG | 9 | **8** | **11** | 18 | 22 | 36 | 17.3 |
| All | 16.0 | 21.2 | **25.0** | **26.8** | 28.8 | 41.2 | 26.5 |

Table 2: Word error rates (percent) for grammar and acoustic constraints. ST-Same Talker, SG-Same Gender, DG-Different Gender. Conditions where our system outperformed human listeners are bolded.

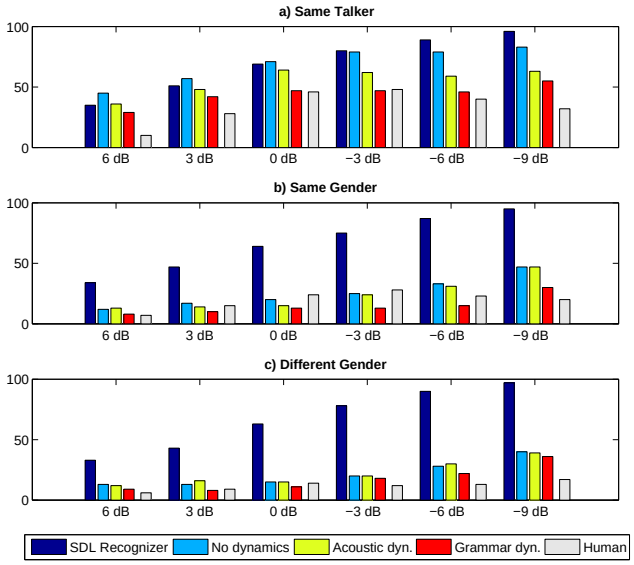



Figure 3: Word error rates for the a) Same Talker, b) Same Gender and c) Different Gender cases.

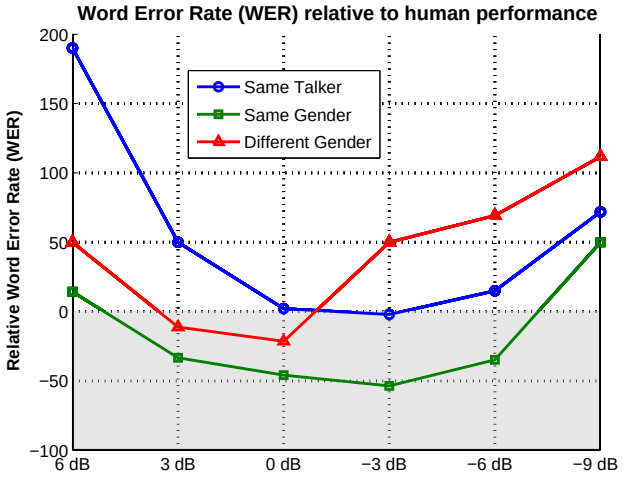



Figure 4: Word error rate of system relative to human performance. Shaded area is where the system outperforms human listeners.

Figure 3 shows results for the 3 different conditions. Human listener performance [6] is shown along with the performance of the SDL recognizer without separation, GMM without dynamics, using acoustic level dynamics, and using both grammar and acoustic-level dynamics.

The top plot in Figure 3 shows word error rates (WER) for the *Same Talker* condition. In this condition, two recordings from the same speaker are mixed together. This conditions best illustrates that acoustic dynamics improved performance considerably relative to no dynamics. Moreover combining grammar and acoustic dynamics improves performance yet again, surpassing human performance in the $-3$ dB condition.

The second plot in Figure 3 shows WER for the Same Gender condition. In this condition, recordings from two different speakers of the same gender are mixed together. In this condition our system surpasses human performance in all conditions except $6$ dB and $-9$ dB.

The third plot in Figure 3 shows WER for the Different Gender condition. In this condition, our system surpasses human performance in the $0$ dB and $3$ dB conditions. Interestingly, temporal constraints do not improve performance relative to GMM without dynamics as dramatically as in the same talker case, which indicates that the characteristics of the two speakers in a short segment are effective for separation.

The performance of our best system, which uses both grammar and acoustic-level dynamics, is summarized in Table 2. This system surpassed human lister performance at SNRs of $0$ dB and $-3$ dB on average across all speaker conditions. Averaging across all SNRs, the system surpassed human performance in the Same Gender condition.

## 7 Discussion

The absolute performance of human listeners is shown in Figure 3. As expected, human listeners perform well when the amplitude of target speaker is considerably higher than the masker. Surprisingly, human listeners also perform well when the target speaker is speaking at a lower amplitude than the masker. Human subjects perform worst when the speakers are at a similar amplitude.

Figure 4 shows the relative Word Error Rate (WER) of our system compared to human subjects. The system performs poorly compared to human subjects when the target speaker is relatively strong. This is to be expected since state of the art ASR systems cannot match human performance for letter recognition. However, the system performs relatively well when the amplitude of the signals is similar. Remarkably, in the *Same Gender* condition, the system is up to 50% better than human subjects. Moreover, when the target speaker is relatively weak, the system does not show the same improvement as human listeners.

It seems that the human auditory system employs different cues and strategies for accomplishing recognition in these conditions. Perhaps human listeners are better able to make use of differences in amplitude as a cue for separation. Further experiments with both human and machine listeners may help to better understand the differences in their performance characteristics.

We are currently investigating which aspects of the system are most important to its performance, and what methods of inference are best for this task. In addition, we hope to overcome some of the limitations of the task. The closed speaker set, artificial mixing conditions, and predictable interference, place limitations on the relevance of the task to realistic scenarios. One of the most important directions for future research will therefore be to address real-world conditions, such as automobile speech recognition with interfering passenger speech.

In conclusion, we have presented some promising models that we feel are just scratching the surface of what is possible. Based on these initial results, we envision that super-human performance over all conditions may be within reach for the speech separation challenge. We also hope that our success in this task will generalize eventually to the unconstrained conditions of everyday life.

## References


[1] T. Kristjansson, J. Hershey, and H. Attias, "Single microphone source separation using high resolution signal reconstruction," *ICASSP*, 2004.

[2] P. Varga and R.K. Moore, "Hidden Markov model decomposition of speech and noise," *ICASSP*, pp. 845–848, 1990.

[3] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, September 1996.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models.," vol. 40, no. 4, pp. 725–735, 1992.

[5] Sam T. Roweis, "One microphone source separation.," in *NIPS*, 2000, pp. 793–799.

[6] Martin Cooke and Tee-Won Lee, "Interspeech speech separation challenge," `http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm`, 2006.

[7] Peder Olsen and Satya Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models," in *Eurospeech 2003*, Geneva, Switzerland, September 1-4 2003, vol. 4, pp. 2509–2512.