INTERSPEECH 2009 BRIGHTON

Constrained Probabilistic Subspace Maps Applied to Speech Enhancement

Kaustubh Kalgaonkar and Mark A. Clements

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta GA USA
{kaustubh, clements}@ece.gatech.edu

10.21437/Interspeech.2009-571

Abstract

This paper presents a probabilistic algorithm that extracts a mapping between two subspaces by representing each subspace as a collection of states. In many cases, the data is a time series with temporal constraints. This paper suggests a method to impose these temporal constraints on the transitions between the states of the subspace.

This probabilistic model has been successfully applied to the problem of speech enhancement and improves the performance of a Wiener filter by providing robust estimates of *a priori* SNR.
Index Terms: speech enhancement, EM, bayesian estimation.

1. Introduction

Additive noise is often present and affects various systems. Noisy speech is not only perceptually difficult to process, it also causes auditory fatigue. Noise also adversely impacts the performance of speech recognition system. To combat these effects, speech enhancement techniques have been employed in a broad range of applications ranging from hearing aids to cellular phones.

Speech enhancement is an important step in many speech related applications, and has been an area of very active research over the years. Various algorithms have been developed to denoise speech depending on the type of noise and application area. Most enhancement algorithms can be broadly classified into Wiener filtering [1], signal subspace modeling [2], statistical and parametric modeling and spectral or cepstral restoration [3] although other methds exist.

One of the most important parameters in many of these algorithms is the a priori signal-to-noise ratio (SNR) estiamte.. Ephraim and Malah [3] proposed a decision-directed (DD) approach to estimate this parameter efficiently.

In this paper we propose a method to probabilistically map two distinct subspaces and use this model to efficiently and reliably estimate the a priori SNR. The a priori SNR estimate using this model improves the performance of a general Wiener filter that uses the DD based approach to estimate a priori SNR and matches the performance of the Ephraim and Malah speech enhancer.

The paper is organized as follows: Section 2 discusses the proposed model, the training algorithm to estimate the model parameters, and the speech production process as it relates to the model. Section 3 explains the speech enhancement problem, the Viterbi solution to find the optimal path and the iterative MAP estimator of the a priori SNR. Section 4 presents the results followed by a small discussion and the conclusion in Section 5 and Section 6 respectively.

2. Probabilistic Subspace Maps

Many signal processing algorithms perform the task of mapping/transforming variables from one subspace (domain) to another (range). This mapping or transform ($\mathbf{T}$) can be *linear* (Kalman filter) or *nonlinear* (Particle Filter/Extended Kalman filter), but in both cases the transform must be known in advance and must be deterministic.

Kalgaonkar and Clements in [4] suggested a method to probabilistically map the domain, range and the transform $\mathbf{T}$, that was modeled as a set of discrete probabilities. The model was successfully applied to perform bandwidth expansion. This paper presents a new model to generate the probability map by imposing constraints on transitions within a subspace. These constraints improve the modeling of temporal characteristics of time-series data such as speech.

A non-constrained subspace model can also be used to perform speech enhancement but, these models do not exploit the process of speech production, and fail to perform better than traditional speech enhancement techniques. The new model derives the constraints on the subspaces from source filter model of speech production.

2.1. The Model

Figure 1 shows the graphical model for the Subspace mapping. The symbols θ and q represent the hidden states that model Subspaces $\mathcal{P}$ and $\mathcal{Q}$ respectively. Given sufficient training data, Subspaces $\mathcal{P}$ and $\mathcal{Q}$ can be modeled with N and M distinct states. The states of the Subspace $\mathcal{Q}$ form a first order hidden Markov chain (HMM) where the transition from a current state q_t to q_{t+1} (t is the time instant) is governed by set of probabilities $\mathbf{A}^q$ such that $p(q_t = j|q_{t+1} = i) = a^q_{ij}$, where $\sum_j a^q_{ij} = 1$. Each of the M states of $\mathcal{Q}$ are modeled with Gaussian mixtures with L Gaussian components as shown in Equation (1)

$$p(\mathbf{x}_t|q_m) = \sum_{l=1}^{L} w^{m,l}\mathcal{N}(\boldsymbol{\mu}_q^{m,l}, \boldsymbol{\sigma}_q^{m,l}) \quad (1)$$

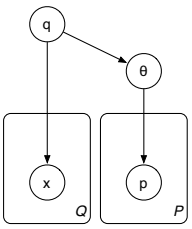


Figure 1: *Graphical Model of the Probabilistic Maps*

Similarly the Subspace $\mathcal{P}$ is mapped with N Gaussians $\mathcal{N}(\boldsymbol{\mu}_\theta^n, \boldsymbol{\sigma}_\theta^n)$ where $n = 1, 2, \ldots, N$. The Gaussians that map both subspaces have diagonal covariances also, the transformation between states of $\mathcal{P}$ and $\mathcal{Q}$ are encoded in a transition matrix $\mathbf{A}$, where $a_{mn} = p(\theta_m|q_n)$ and $\sum_m a_{mn} = 1$

2.2. Parameter Estimation using EM

Each subspace of the model has a different structure, and different set of parameters. The Subspace $\mathcal{Q}$ is modeled with a HMM which has three parameters to estimate, $\Lambda = \{\pi, \mathbf{A}^q, p(\mathbf{x}_t|q_m)\}$, where π are the initial probabilities of the states. Estimation of these parameters is performed using standard expectation maximization (EM) [5], that involves two steps:

Compute the a posteriori probabilities during the E-step:

$$p(\mathbf{x}_t, \mathbf{p}_t|q_m, \theta_n) = \frac{p(\mathbf{x}_t, \mathbf{p}_t, q_m, \theta_n)}{\sum_{m=1}^{M}\sum_{n=1}^{N} p(\mathbf{x}_t, \mathbf{p}_t, q_m, \theta_n)} \quad (2)$$

where the joint probability is given by Equation (3)

$$p(\mathbf{x}_t, \mathbf{p}_t, q_m, \theta_n) = p(\mathbf{p}_t|\theta_n)p(\theta_n|q_m)p(\mathbf{x}_{(1:t)}, q_m) \quad (3)$$

As the Subspace $\mathcal{Q}$ is modeling a time series the term $p(\mathbf{x}_{(1:t)}, q_m)$ is the probability of observing a sequence of $\mathbf{x}$ upto time instant t, this is computed using the forward-backward algorithm for HMM [6].

In the M-step, the complete data likelihood $\mathcal{L}$ is maximized:

$$\mathcal{L} = \mathbf{E}_{q,\theta|\mathbf{x},\mathbf{p},\Omega}\{\log p(\mathbf{x}_t, \mathbf{p}_t, q_m, \theta_n)\} \quad (4)$$

where $\Omega = \{\Lambda, \mathbf{A}, \Theta\}$, are the complete model parameters.

Parameter estimation is performed by alternatively solving Equations (3) and (4). Details of the parameter estimation will not be presented here due to space constraints.

Solving for the parameter of Subspace $\mathcal{Q}$

$$\pi_m = \frac{\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)}{\sum_{i=1}^{M}\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)} \quad (5)$$

$$a^q_{ms} = \frac{\sum_{t=1}^{T-1}\sum_{n=1}^{N} p(q_m, q_s, \theta_n|\mathbf{x}_t, \mathbf{p}_t)}{\sum_{t=1}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)} \quad (6)$$

$$w^{m,l} = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q^l_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)}{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)} \quad (7)$$

$$\boldsymbol{\mu}^{m,l}_q = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q^l_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)\mathbf{x}_t}{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)} \quad (8)$$

$$\boldsymbol{\sigma}^{m,l}_q = \frac{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q^l_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)(\mathbf{x}_t - \boldsymbol{\mu}^{m,l}_q)^2}{\sum_{t=1}^{T}\sum_{n=1}^{N} p(q_m, \theta_n|\mathbf{x}_t, \mathbf{p}_t)} \quad (9)$$

The parameters of the HMM can be efficiently computed using Equations (5) through (9) and a modified version of Baum-Welch method [6].

Parameters of Subspace $\mathcal{P}$, $\Theta = \{\boldsymbol{\mu}_\theta, \boldsymbol{\sigma}_\theta\}$ can be similarly estimated.

2.3. Speech Production and Probabilistic Subspace Maps

The source filter model breaks speech production into two blocks: the glottal source which provides the excitation and the Vocal Tract (VT) which shapes the excitation to produce speech. It is impossible to know the real excitation and VT characteristics even while measuring the pressure or velocity at the lips of the talker. Without actual measurements, it is possible to make multiple indirect observations of both the vocal tract and excitation. The speech spectrum is one such observation.

The physiology of the talker who performs at finite speed, and has a vocal tract with finite flexibility, imposes constraints on the vocal tract time series that will be used to model the Subspace $\mathcal{Q}$. We model the actual vocal tract as the hidden state in Subspace $\mathcal{Q}$. The anatomical constraints put physical restrictions on transition between the states 'q_m'of the Subspace $\mathcal{Q}$. Linear Prediction Coefficients, Partial Correlation coefficients, Log Area Ratios and Line Spectral Pairs (LSP) all provide valid observations that can help infer the true state of Subspace $\mathcal{Q}$, the actual vocal tract shape. In these experiments LSP along with the gain of the predictor polynomial were used as the observations $\mathbf{x}_t$.

One of the interesting properties about the vocal tract area function is its many-to-one relationship with the speech spectrum, as shown by Kalgaonkar and Clements [7]. This multiplicity explains the many-to-one functional mapping between Subspace $\mathcal{Q}$ and Subspace $\mathcal{P}$ which, probabilistically models the magnitude spectra of speech. This simplified model is shown in Figure 2. The shaded circles represent observations of the subspaces.

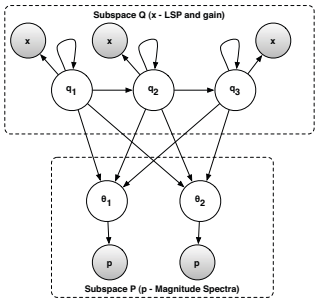


Figure 2: *Probabilistic Subspace Map for Speech Enhancement*

3. Speech Enhancement

Let s and n denote the speech and the uncorrelated additive noise signals and let $y = s + n$ be the observed signal.Then the speech enhancement problem can be formulated as finding the estimator that minimizes the conditional expectation of distortion($\mathcal{D}$) given the noisy speech.

$$\hat{s} = \underset{\hat{s}}{\operatorname{argmin}}\ \mathbf{E}\{\mathcal{D}(s, \hat{s})|y\} \quad (10)$$

The problem of enhancement can be efficiently performed in the spectral domain. Applying a short-time Fourier Transform to the observed signal, the problem now becomes $Y(k,l) = X(k,l) + N(k,l)$, where $Y(k,l), X(k,l)$ and $N(k,l)$ are the magnitude spectra of the noisy speech, clean speech and noise, for the k^{th} frame of speech and the l^{th} frequency bin. The indices k and l will be dropped for brevity.

The generalized solution of the problem (10) can be written as $\widehat{X} = G(\xi, \eta)Y$, where G is the gain of the denoising filter which is a function of $\xi = (\frac{X}{N})^2$ the a prior SNR and $\eta = (\frac{Y}{N})^2$ the a posteriori SNR. This convention is adopted from McAulay et al. [8]

The specifics of the gain function depend on the choice of distortion measure, for a Wiener filter that uses squared-error as a

distortion measure, the gain function is given by:

$$G_W(\xi) = \frac{\xi}{1+\xi} \tag{11}$$

As seen from the Equation (11), the goal of Wiener filtering is to estimate the a priori SNR (true power spectrum) given the noisy measurements, which is not an easy task. One way to circumvent the problem is to estimate the noise spectrum, and compute the Weiner filter using $G = \frac{Y^2 - N^2}{Y^2}$, or use the decision directed approach to compute the a priori SNR.

We solve the problem of estimating the a priori SNR by using the constraint maps explained in Section 2.3. The estimation problem now becomes:

$$\max_{\mathbf{x}_t, q_{(1:t)}} p(\mathbf{x}_t, q_t|\mathbf{p}_{(1:t)}) = \max_{\mathbf{x}_t} p(\mathbf{x}_t|q_t) \max_{\mathbf{q}_{(1:t)}} p(q_{(1:t)}|\mathbf{p}_{(1:t)}) \tag{12}$$

The problem in Equation (12) involves two steps. The first step estimates the value of $\mathbf{x}_t$ given an optimum/most likely state q_m^* at time t and the second step of the problem is to estimate the most likely state sequence $q^*_{(1:t)}$ given the spectral observations $\mathbf{p}_{(1:t)}$. The two problems are separable and can be independently solved in reverse order.

3.1. Finding the Most Likely State Sequence

The problem of estimating the optimal state sequence can be rewritten as:

$$\max_{q_{(1:t)}} p(q_{(1:t)}|\mathbf{p}_{(1:t)}) \max_{q_{(1:t)}} \sum_{n=1}^{N} p(q_{(1:t)}, \mathbf{p}_{(1:t)}, \theta_n) \tag{13}$$

$$= \Big[\max_{q_{(t)}} p(q_t|q_{t-1}) \sum_{n=1}^{N} p(q_t, \mathbf{p}_t, \theta_n)\Big] \max_{q_{(1:t-1)}} p(q_{(1:t-1)}|\mathbf{p}_{(1:t-1)}) \tag{14}$$

Equation (14) is solved using the Viterbi algorithm, to yield the most likely path $q^*_{(1:t)}$ in Subspace $\mathcal{Q}$.

3.2. Estimating the Speech Spectrum

Given the most likely path through the Subspace $\mathcal{Q}$, the task now boils down to estimating the vocal tract parameters $(\mathbf{x}_t)$ and the a priori SNR. The problem of estimating $\mathbf{x}_t$ given the optimal path can also be framed as:

$$\max_{\mathbf{x}_t} \sum_{q_t} p(q_t|q^*_{(t-1)}) p(\mathbf{x}_t|q_t) \Phi(q_t) \tag{15}$$

where $\Phi(q_t)$ is the $\sum_{n=1}^{N} p(\mathbf{p}_t|\theta_n) p(\theta_n|q_t)$. Maximizing (15) with respect to $\mathbf{x}_t$ gives a simple iteration (16), that generally converges in 3 to 5 steps.

$$\mathbf{x}_t^{(k)} = \frac{\sum_{m=1}^{M} p(q_t|q^*_{(t-1)}) \Phi(q_t) \sum_{l=1}^{L} w^{m,l} \mathcal{N}^{m,l}(\mathbf{x}_t^{(k-1)}) (\frac{\boldsymbol{\mu}_q^{m,l}}{(\boldsymbol{\sigma}_q)^{m,l}})}{\sum_{m=1}^{M} p(q_t|q^*_{(t-1)}) \Phi(q_t) \sum_{l=1}^{L} w^{m,l} \mathcal{N}^{m,l}(\mathbf{x}_t^{(k-1)}) (\frac{1}{(\boldsymbol{\sigma}_q)^{m,l}})} \tag{16}$$

where 'k' indicates the iteration number.

The a priori spectrum of speech can be estimated from the LSP coefficients and the gain using (17)

$$\widehat{X} = \frac{g^2}{P_{ss}(f)} \tag{17}$$

where $P_{ss} = (|A(\exp j2\pi f/f_s)|)^2$, A is the prediction polynomial estimated form $\mathbf{x}_t$, and g is the linear prediction gain.

In the presence of noise, the speech signal no longer resides in the Subspace $\mathcal{P}$ but is now present in a modified Subspace $\overline{\mathcal{P}}$. The parameters of this new Subspace $\overline{\mathcal{P}}$ can be derived given the noise model. As an example, if the noise n corrupting the speech s is White and is modeled with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$, the parameters of subspace $\overline{\mathcal{P}}$ are given by Equation (18) and (19)

$$\boldsymbol{\mu}_\theta^n = \boldsymbol{\mu}_\theta^n + \delta\boldsymbol{\mu} \tag{18}$$

$$\boldsymbol{\sigma}_\theta^n = \boldsymbol{\sigma}_\theta^n + \delta^2\boldsymbol{\sigma} \quad where \;\; 1 \leq n \leq N \tag{19}$$

Where δ is the variation in the energy between training and estimated noise.

The complete speech enhancement procedure is given in Algorithm 1.

Algorithm 1 Denoise Speech using the Probabilistic mapping

1: **while** Noisy Speech **do**
2: Estimate the noise spectrum.
3: Update the noise gain δ.
4: Adapt Subspace $\mathcal{P}$ using Equations (18) and (19).
5: Estimate the ML state sequence $q^*_{(1:t)}$ using Equation (14).
6: Estimate the the VT parameters using Equation (16).
7: Compute the a priori SNR using Equations (17) and noise estimate from Step 2.
8: Compute the Wiener gain using Equation (11).
9: Estimate the clean speech using the Wiener filter generated in Step 8
10: **end while**

4. Experiments and Results

Experiments were conducted using recordings from six speakers, three males and three females. The recordings were obtained from the Wall Street Journal Database, using 15-20 minutes of 16KHz data from each speaker. Five type of noise recordings 'babble', 'Pink', 'Volvo', 'White' and 'Factory' were selected from the NOISEX-92 database. Speech enhancement tests were conducted on audio data not used in training the model.

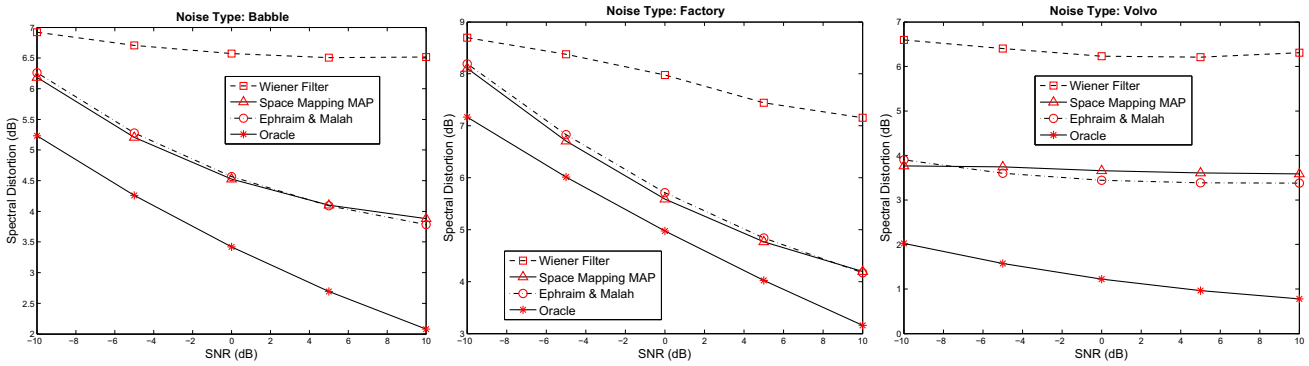
Both training and test data were analyzed with a 32 msec Hamming window with 50% overlap between adjacent frames. 16 LSP coefficients were obtained for each frame, which form the observations of Subspace $\mathcal{Q}$. A 512 point FFT was performed for each frame to obtain 257 distinct coefficients for Subspace $\mathcal{P}$. Models of various sizes (M and N) were trained, but the results presented here are for a model with $M = 75$ and $N = 300$. In our experiments we found that this model size was sufficient for the number of speakers that were involved in the experiments. Reduction in the size of model to $(M, N) = (40, 100)$ degraded the performance of the system by 0.5 to 1 dB.

A simple three-class, voice activity detector (VAD) based on spectral distance, zero crossing rate, and energy was used to isolate noisy speech frames from noise. Noise was modeled with a 10 component Gaussian mixture model prior to the start of the enhancement procedure. This noise model was continuously updated with the information from the VAD. We used the same noise estimator for the evaluation of the performance of all the enhancement algorithms.

The performance of the system was measured in terms of spectral distortion given by:

$$D = \frac{20}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \Big| \Big(\log_{10} |X(k,l)| - \log_{10} |\widehat{X}(k,l)| \Big) \Big| \tag{20}$$

Figure 3 shows comparative performance for four speech enhancement techniques: Wiener Filter (WF) with decision directed a prior SNR estimation, Wiener filter with a prior SNR estimated using constrained Probabilistic maps (WPM), Ephraim and Malah speech enhancer (E&M) [3] and an oracle Wiener filter with perfect knowledge of the speech and noise Spectra (OWF). As seen



(a) Babble Noise (b) Factory Noise (c) Volvo car Noise

Figure 3: Spectral Distortion vs SNR comparison plots for Wiener, Wiener modified with Probabilistic Maps, Ephraim and Malah and Oracle Wiener (Both noise and speech spectra known)

in the figures, the performance of the WPM is very similar to the performance of the E&M algorithm.

The use of constrained probabilistic maps to estimate the speech spectrum improves the estimation of the a priori SNR by providing a better and more reliable estimate of the speech spectra, which consequently improves the spectral distortion performance of WPM over the normal WF by 2 dB on average. Table 1 lists the spectral distortion performance for Pink noise.

Table 1: *Spectral Distortion for Pink Noise*

SNR(dB)	Spectral Distortion (dB)			
	WF	WPM	E&M	OWF
-10	8.3219	8.3787	8.1083	7.8662
-5	8.1896	6.8669	6.7107	6.7076
0	8.1264	5.6072	5.5764	5.6406
5	7.7252	4.6699	4.7029	4.6440
10	7.4998	4.0825	4.1229	3.7189

5. Discussion

The purpose of this paper is to suggest a framework for mapping subspaces using probabilities. Speech enhancement is an interesting application of this framework. One of the important distinctions between the proposed enhancement algorithm and the existing technique is the lack of parameter modeling. The current technique does not suggest a model of LPC or spectra but establishes a mapping which in the real world is nonlinear and difficult to extract. Further the impact of additive noise on this mapping cannot be clearly extracted otherwise.

This mapping is now exploited to estimate clean speech spectrum from noisy measurements. E&M used elaborate statistical models for speech and noise. In this work no such models were used to improve the performance of the Wiener filter to parallel that of E&M. Our goal is to exploit noise statistics to jointly estimate the clean speech and noise spectrum thereby providing a Wiener filter that approaches performance of the Oracle.

6. Conclusions and Future work

This paper presents a new statistical model that probabilistically maps subspaces and transforms between subspaces. This model also imposes constraints on state transition within a subspace by using a HMM to model the subspace.

This new model is applied to the problem of speech enhancement. The algorithm suggested in the paper improves the performance of the general Wiener filter. Informal listening tests have shown that the perceptual performance of the WPM Wiener filter is better than the rest of the algorithms This performance boost is achieved by improving the estimate of the a priori SNR, that is used to estimate the filter gain.

This paper presents our preliminary work on application of the model for speech enhancement. The target of this research is to get the performance of the Wiener filter close to the Oracle, by exploiting the characteristics and non stationarity of the noise. We are also exploring the application of the model to other problems like speaker separation.

7. References

[1] J.H.L. Hansen and M.A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *Signal Processing, IEEE Transactions on*, vol. 39, no. 4, pp. 795–805, Apr 1991.

[2] A. Rezayee and S. Gazor, "An adaptive klt approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 2, pp. 87–95, Feb 2001.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[4] K. Kalgaonkar and M. A. Clements, "Sparse probabilistic space mapping and it application to speech bandwidth expansion," in *ICASSP*, 2009.

[5] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[6] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, August 2006.

[7] K. Kalgaonkar and M. A. Clements, "Vocal tract and area function estimation with both lip and glottal losses," in *Interspeech, Antwerp Belgium*, 2007, pp. 550–553.

[8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 2, pp. 137–145, Apr 1980.