SPARSE PROBABILISTIC STATE MAPPING AND ITS APPLICATION TO SPEECH BANDWIDTH EXPANSION

Kaustubh Kalgaonkar & Mark A Clements

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
{kaustubh,clements}@ece.gatech.edu

ABSTRACT

In this paper we present a probabilistic algorithm that extracts a mapping between two subspaces by representing each subspace as a collection of states. An arbitrary increase in number of states results in over-fitting the training data without exploring the underlying structure of the map. This paper suggests a method to impose sparsity constraints on the state map by using entropic priors.

This probabilistic model is applied to the problem of artificial bandwidth expansion that involves estimating the missing frequency components (3.7 – 8 kHz and 0 – 0.3 kHz) of speech given the narrowband speech signal (0.3 – 3.7 kHz).

Index Terms— Bandwidth expansion, Signal reconstruction, Sparse representation.

1. INTRODUCTION

Artificial Bandwidth Expansion (ABE) is a process of automated addition for missing high frequency components to a bandlimited speech signal. Various techniques have been proposed for this task over the years. Listening tests have shown that the presence of high frequency components in speech make it perceptually more pleasing thereby improving its perceived quality [1]. Most of the techniques focus on extending the bandwidth of a telephony (300Hz to 3700 Hz) signal producing a speech signal in the range 0Hz to 8000 Hz. Artifact-free synthesized speech is one of the pivotal requirements of a good ABE system.

Aliasing-based methods (e.g., [2]) employ a non-linear transformation to construct the absent high frequency components by aliasing low frequency components. Some methods [3, 4] use codebooks to generate a map between the low and the missing high frequencies of the spectrum. Methods such as [5] attempt to estimate missing spectral components as a linear combination of the low frequency components.

Statistical methods such as those proposed in [6, 7] model the relationship between the lower and upper band frequency components using Hidden Markov Models (HMM), Gaussian Mixture Models (GMM) etc. The trained statistical models are then used to estimate the missing frequency components.

In our previous work [6, 8], we have used Vocal Tract (VT) areas to perform ABE. This method does not directly estimate the missing spectral components, but rather estimates the vocal tract area function necessary for the production of the broadband speech. This method uses a combination of codebook and statistical methods of bandwidth extension.

In this paper we propose a method based on probabilistic mapping of subspaces that takes advantage of the inherent sparsity in the system to generate an over-complete bases of the target subspace, which can then be used to produce a MMSE estimate of the missing frequency components.

The paper is organized as follows: Section 2 explains the estimation problem and the probabilistic space mapping algorithm. Sections 3 presents the application of this algorithm to bandwidth expansion. Section 4 presents the experiments used to evaluate the performance of the system and the results. Section 5 presents the conclusion and discusses the future direction of research.

2. PROBABILISTIC STATE MAPPING

A host of signal processing applications involve mapping or estimating $\mathbf{q} \in \mathcal{Q}$ from $\mathbf{p} \in \mathcal{P}$ where $\mathcal{P}$ and $\mathcal{Q}$ might be the same subspace. Applications such as speech modification, denoising and speaker separation fall in to this category. Kalman filter and particle filter are some of the methods developed to perform these tasks. Knowledge of this mapping function 'f' is one of the fundamental requirements of these traditional methods.

In this paper we propose a graphical system that will probabilistically model the mapping between the subspaces $\mathcal{P}$ and $\mathcal{Q}$ given sufficient training data. Figure 1 shows the graphical model that is used to represent this state mapping. π and γ are hidden variables that model the subspaces $\mathcal{P}$ and $\mathcal{Q}$ respectively. Assume that subspaces $\mathcal{P}$ and $\mathcal{Q}$ can be modeled with N and M distinct bases. Assume that the subspace $\mathcal{P}$ is modeled by N Gaussians $\mathcal{N}(\boldsymbol{\mu}_\pi^n, \boldsymbol{\sigma}_\pi^n)$, where $n = 1, 2, \ldots, N$ and the subspace $\mathcal{Q}$ is modeled with M Gaussians $\mathcal{N}(\boldsymbol{\mu}_\gamma^m, \boldsymbol{\sigma}_\gamma^m)$ where $m = 1, 2, \ldots, M$.

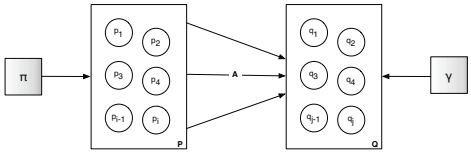


Fig. 1. Graphical model representing the mapping between states of subspaces $\mathcal{P}$ and $\mathcal{Q}$.

The relation between the states of $\mathcal{P}$ and $\mathcal{Q}$ is encoded in the transition matrix $\boldsymbol{A}$ where $a_{mn} = p(\gamma_m|\pi_n)$ and $\sum_m a_{mn} = 1$. The joint probability $p(\mathbf{p}_t, \mathbf{q}_t, \pi_n, \gamma_m)$ can be written as:

$$p(\mathbf{p}_t, \mathbf{q}_t, \pi_n, \gamma_m) = p(\mathbf{q}_t|\gamma_m)p(\gamma_m|\pi_n)p(\mathbf{p}_t|\pi_n)p(\pi_n) \quad (1)$$

The best place to motivate model parameter estimation is to start with the solution of the estimation problem. Given the trained model

$\mathcal{M} = \{\mathbf{\Pi}, \mathbf{\Gamma}, \boldsymbol{A}\}$ and $\mathbf{p}_t$, the minimum mean square error (MMSE) estimate of $\widetilde{\mathbf{q}}_t$ is given by Equation (2)

$$\widetilde{\mathbf{q}}_t = \mathbf{E}_{\mathbf{q}|\mathbf{p}}\{\mathbf{q}|\mathbf{p}\} = \int \mathbf{q} \cdot p(\mathbf{q}|\mathbf{p}) d\mathbf{q} \tag{2}$$

The conditional probability $p(\mathbf{q}|\mathbf{p})$ can be expressed as the marginal of the joint probability (1)

$$p(\mathbf{q}|\mathbf{p}) = \frac{p(\mathbf{p}\ \mathbf{q})}{p(\mathbf{p})} = \frac{\sum_{m=1}^{M}\sum_{n}^{N} p(\mathbf{p}, \mathbf{q}, \pi_n, \gamma_m)}{\sum_{n=1}^{N} p(\mathbf{p}|\pi_n)p(\pi_n)} \tag{3}$$

Using Equations (3), (2) and (1) the MMSE estimate of $\widetilde{\mathbf{q}}_t$ can be written as

$$\widetilde{\mathbf{q}}_t = \sum_{m=1}^{M} \boldsymbol{\mu}_{\gamma}^{m} \left[\frac{\sum_{n}^{N} p(\gamma_m|\pi_i)p(\mathbf{p}_t|\pi_n)p(\pi_n)}{\sum_{n=1}^{N} p(\mathbf{p}|\pi_n)p(\pi_n)} \right] \tag{4}$$

$$\widetilde{\mathbf{q}}_t = \mathbb{M}_{\gamma}\boldsymbol{\alpha}_t \tag{5}$$

where $\mathbb{M}_{\gamma} = [\boldsymbol{\mu}_{\gamma}^{1} \vdots \boldsymbol{\mu}_{\gamma}^{2} \vdots \cdots \vdots \boldsymbol{\mu}_{\gamma}^{M}]$ is the matrix of bases formed by the means of the Gaussian mapping the subspace $\mathcal{Q}$ and $\sum_m \alpha_m = 1$ is the matrix of probabilities containing the belief for each basis. According to the Equation (5) the estimate $\widetilde{\mathbf{q}}$ is the *convex sum* of the bases mapping the subspace.

The MMSE estimates obtained with this model will lie in the *convex hull* of the bases vectors, any point outside the hull is estimated with error. The performance of the estimator depends on the placement of these bases vectors and the resulting convex hull. The model is trained by making an educated guess on the number of bases that would be required to map the subspaces $\mathcal{P}$ and $\mathcal{Q}$, over-estimating the number of bases will lead to over-fitting the model to available training data. These over-fitted bases will fail to extract the underlying mapping and structure of the data. Figure 2(a) shows the six bases of the subspace $\mathcal{Q}$ estimated without any sparsity constraints. On the other hand over-fitted bases with sparsity constraints on the transition matrix $\boldsymbol{A}$ will permit a flexible mapping and hence a larger convex hull. In the next section we will describe the constraints that can be applied to the model to generate set of over-complete bases better representing the structure of data.

2.1. Sparsity constrained probabilistic mapping

Under the current model every state π_n from subspace $\mathcal{P}$ maps to a state γ_m in subspace $\mathcal{Q}$ with a probability a_{mn}. In an over-complete bases case (where number of bases exceeds the the dimension of the space) this might be completely unnecessary. One particular instance of $\mathbf{q}$ could be completely described by only a subset of the bases, implying each input state π_n only maps to a handful of the output states γ_m. This sparsity constraints on columns of $\boldsymbol{A}$ has to be imposed while training the model.

Various metrics have been applied to measure and impose sparsity, L_p norms are one of the most popular measures of sparsity [9].

This paper imposes the sparsity using *entropic prior*. Given a probability distribution θ we can write the entropic prior for the distribution as $P_e(\theta) \propto \exp(-\beta\mathcal{H}(\theta))$ for a multinomial distribution entropy $\mathcal{H}(\theta) = -\sum_i \theta_i \log\theta_i$. Positive values of sparsity parameter β favor distributions with lower entropy [10]. The distribution θ corresponds to the $p(\boldsymbol{\pi})$ and $p(\gamma_m|\pi_n)$ for $n = 1, 2, \ldots, N$. Entropic priors will be use to trim the excess states in subspace $\mathcal{P}$ and impose sparsity on the columns of transition matrix $\boldsymbol{A}$.

2.2. Parameter estimation

Parameter estimation is performed using the a small variation to the EM algorithm. Parameters for the Gaussians $\boldsymbol{\mu}_{\gamma,\pi}, \boldsymbol{\lambda}_{\gamma,\pi}$ are estimated using the traditional EM, $p(\pi)$ and $\mathbf{p}(\gamma|\pi)$ are estimated using *maximum a posterior* (MAP) estimation with *entropic priors*

A posteriori portability is computed for the E-step:

$$p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t) = \frac{p(\mathbf{p}_t, \mathbf{q}_t, \pi_n, \gamma_m)}{\sum_{n=1}^{N}\sum_{m=1}^{M} p(\mathbf{p}_t, \mathbf{q}_t, \pi_n, \gamma_m)} \tag{6}$$

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

$$\mathcal{L} = \mathbf{E}_{\gamma,\pi|\mathbf{pq},\mathcal{M}}\{\log p(\mathbf{p}_t, \mathbf{q}_t, \pi_n, \gamma_m)\} \tag{7}$$

Solving the Equation (7) for the parameters of the Gaussians (mean and variance) in subspaces $\mathcal{P}$ yields:

$$\boldsymbol{\mu}_{\pi}^{n} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)\ \mathbf{p}_t}{\sum_{t=1}^{T}\sum_{m=1}^{M} p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)} \tag{8}$$

$$(\boldsymbol{\sigma}^2)_{\pi}^{n} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)\ (\mathbf{p}_t - \boldsymbol{\mu}_{\pi}^{n})^2}{\sum_{t=1}^{T}\sum_{m=1}^{M} p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)} \tag{9}$$

The parameters of Gaussians of subspace $\mathcal{Q}$ can be written similarly to those of subspace $\mathcal{P}$ and are omitted here due to the space constraints.

Entropic estimation of $p(\boldsymbol{\pi})$ is performed by maximizing the new augmented likelihood $\mathcal{R}$

$$\mathcal{R} = \mathcal{L} + \tau\Big(\sum_{n} p(\pi_n) - 1\Big) + \delta\sum_{n} p(\pi_n)\log p(\pi_n) \tag{10}$$

where τ is the Lagrange multiplier and δ is the parameter that controls the sparsity. The M-step for estimating $p(\pi_n)$ reduces to:

$$\frac{\omega_n}{p(\pi_n)} + \delta + \delta\log p(\pi_n) + \tau = 0 \tag{11}$$

where ω_n represents the expected sufficient statistice in this case $\sum_t\sum_m p(\pi_n, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)$. Equation (11) is a system of simultaneous transcendental equations and can be solved using the Lambert $\mathcal{W}$ function [11] as suggested by Brand [10] to yield:

$$p(\pi_n) = \frac{-\omega_n/\delta}{\mathcal{W}(-\omega_n e^{1+\tau/\delta}/\delta)} \tag{12}$$

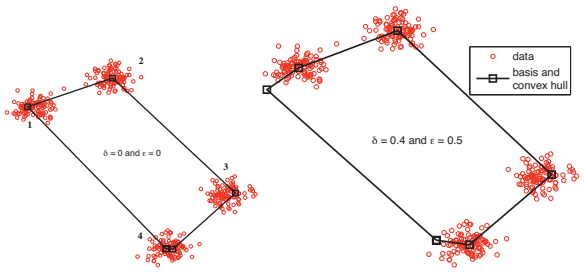
Equations (11) and (12) form a pair for fixed-point iteration that typically converges in 2-5 iterations.

Similar technique can be used to obtain the columns of the transition matrix $\boldsymbol{A}$, where $\omega_m^j = \sum_t p(\pi_j, \gamma_m|\mathbf{p}_t, \mathbf{q}_t)$ will estimate a sparse representation of the j^{th} column of the transition matrix. The Figure 3 shows the effect of sparsity constraints on the transition matrix.

Final model update equations are given by (6), mean and variance updates for Gaussians in both $\mathcal{P}$ and $\mathcal{Q}$ subspaces using (8) and (9), and updates of transition matrix $\boldsymbol{A}$ and $p(\boldsymbol{\pi})$ using fixed-point update equations (11) and (12).

2.3. Effect of sparsity constraints on the model.

The following synthetic data provides a good opportunity to illustrate the difference between complete and sparse parameterization of a probabilistic map. Figure 2 shows four distinct data clusters that form subspace $\mathcal{Q}$. Subspace $\mathcal{P}$ is formed by taking the projection of the data on the x-axis. Also the x and y dimensions of the clustered data are uncorrelated. Most of the traditional algorithms will have difficulty in estimating the $\mathbf{q}$ given $\mathbf{p}$.

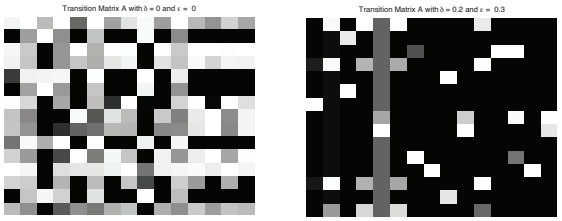


(a) Over-complete bases. (b) Over-complete bases with sparsity constraints.

Fig. 2. Over-complete bases vectors extracted from the data and the convex hull.

Two sets of six over-complete bases are obtained using the EM algorithm. The first set of bases was obtained without any constraints on $p(\boldsymbol{\pi})$ and $\boldsymbol{A}$. This basis set and its convex hull are shown in Figure 2(a). Without any constraints, the multiple basis/Gaussians were placed near the means of the Clusters 1 and 4. This model ignores the overlapping x-projections of Clusters 2 and 4 and overlapping y-projections of Clusters 1 and 2. The convex hull formed with these bases is around the means of the four clusters.

The second set of bases were obtained by imposing sparsity $\epsilon = 0.5$ on the transition matrix $\boldsymbol{A}$ and $\delta = 0.4$ on $p(\boldsymbol{\pi})$. Two bases in this case have moved away from the means of Clusters 1 and 4, forming a larger convex hull, as seen in the Figure 2(b). Additional bases near Clusters 1 and 4 will be able to provide additional resolution when reconstructing data points of Clusters 4 and 2 thereby reducing the MSE in reconstructing y from x.



(a) No Sparsity imposed. (b) With sparsity constraints $\delta = 0.2$ and $\epsilon = 0.3$.

Fig. 3. Transition Matrix $\boldsymbol{A}$.

3. APPLICATION: ARTIFICIAL BANDWIDTH EXTENSION

The ABE is performed in the spectral domain. Figure 4 shows the block diagram of the ABE system. There are two stages to the ABE

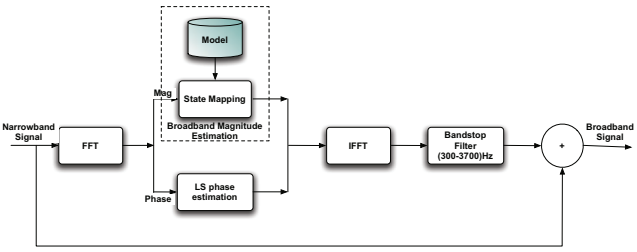


Fig. 4. Block diagram of ABE system.

system: *magnitude estimation* and *phase estimation*, both the stages are explained in detail in the next subsections. The synthesized broadband speech is then passed through a bandstop filter to only retain the missing components which are then added to the original narrowband signal.

3.1. Magnitude Estimation

Let $\mathbf{p} \in \mathbb{R}^k$ be the magnitude spectrum of the narrowband signal and $\mathbf{q} \in \mathbb{R}^l$ be the magnitude spectrum of the broadband signal where $k \leq l$. Probabilistic mapping $\mathcal{M}$ between the broadband and the narrowband spectra is obtained using the method described in Section 2.1.

The trained model $\mathcal{M}$ is used to estimate the high-frequency components of the broadband signal from the narrowband speech using Equation (5).

3.2. Phase Estimation

The naive approach of using the phase of the low-frequency components to synthesize the broadband speech will produce artifacts in the synthesized broadband audio. A better approach is to use a simple linear transform $\boldsymbol{T}$ to estimate the phase of the broadband signal $\phi_{\mathbf{q}}$ using the phase of the narrowband signal $\phi_{\mathbf{p}}$.

The transform matrix can be learned from training data using a simple LLSE given by:

$$\boldsymbol{T} = \Phi_{\mathbf{q}} \Phi_{\mathbf{p}}^{\dagger} \tag{13}$$

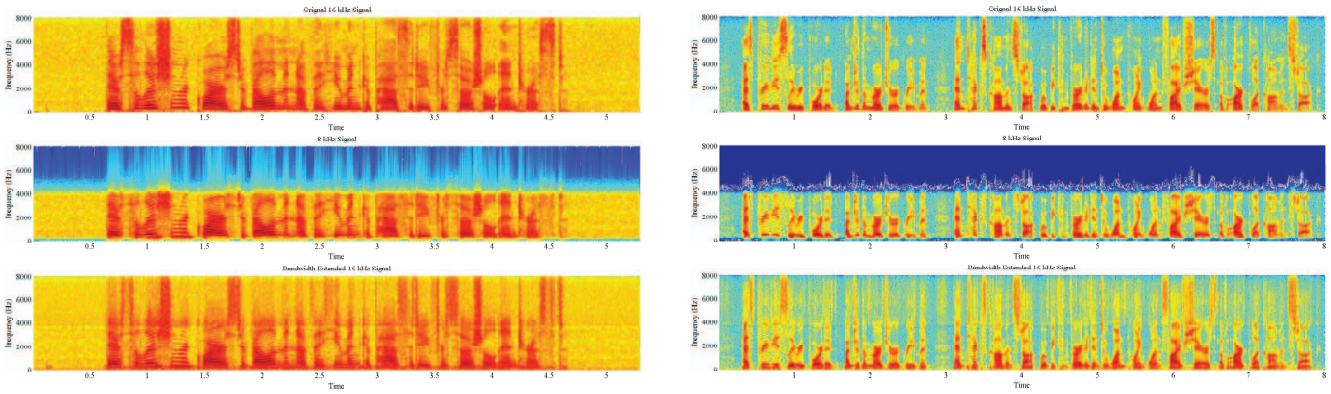
where $\Phi_{\mathbf{p}}, \Phi_{\mathbf{q}}$ are the matrices of phases of the narrowband and broadband speech and $\dagger$ is the pseudoinverse operation.

4. EXPERIMENTS AND RESULTS

Experiments were conducted on recordings from six speakers three males and three females. The recordings were obtained from the Wall Street Journal database, using 15–20 min of high bandwidth data sampled at 16 kHz. The narrowband speech was obtained by bandpass filtering the broadband speech with a 8^{th} order Chebyshev filter with cutoff at 300 and 3700 Hz. Both training and testing data were analyzed with a 32ms window with 50% overlap between the adjacent frames. The signal was windowed using a Hanning window. The spectrum was obtained using 512 point FFT resulting in magnitude spectrum with 257 unique points.

Testing was performed on the data that disjoint the training set. Equation (14) is used to measure the spectral distortion between the reconstructed $\widehat{P}_{ss}(f)$ and orignal $P_{ss}(f)$ high bandwidth signal.

$$D^2 = \frac{1}{f_s} \int_0^{f_s} \left(20 \log_{10}(P_{ss}(f)) - 20 \log_{10}(\widehat{P}_{ss}(f))\right) df \tag{14}$$



(a) Female Speaker. (b) Male Speaker.

Fig. 5. Spectrogram for speech data for two speakers. From top original broadband speech sampled at 16 kHz, narrowband 8 kHz speech and 16 kHz bandwidth expanded speech reconstruction was performed using 512 bases with $\delta = 0.1$ and $\epsilon = 0.4$ and frame size of 32 ms.

where f_s is the sampling frequency in Hz and A is the linear prediction polynomial and $P_{ss}(f) = (|A(\exp(\jmath 2\pi f/f_s))|)^{-1}$.

Experiments were performed on data by varying window sizes, number of bases and the sparsity parameters δ for $p(\boldsymbol{\pi})$ and ϵ for the transition matrix. Table 1 shows the spectral distortion with respect to number of bases for a frame size of 128 samples. The models used for these experiments were trained independent of the speaker. Speaker dependent models display better performance. We observed 0.5 dB improvement in spectral distortion when the ABE was performed using speaker dependent models with 80 bases and a frame size of 128 samples. We observed an improvement in performance

Table 1. Number of bases and Spectral Distortion

bases	20	40	64	80	80 ($\epsilon = 0.3$)
$D^2(dB)$	4.5824	4.3329	4.1733	4.0683	4.0024

with increase in the frame size. The increase in the size of the FFT requires an increase in the number of bases to maintain the performance. This algorithm improves upon the performance of our previous ABE system [6] and within the measure, this algorithm outperforms the system suggested by [12]. Figure 5 shows spectrograms for a male and a female speaker. In both cases the algorithm is able to reconstruct the missing frequencies both in the 0 – 300 Hz and 3700 – 8000 Hz region artifact-free. Examples of the reconstructed speech can be found at: `www.ece.gatech.edu/~kaustubh/bandexp/icassp09.html`.

5. CONCLUSION

In this paper we have presented a probabilistic subspace mapping algorithm that exploits the inherent sparsity between the mapping of the subspaces to improve the MMSE estimate obtained by the model. Proper choice of the sparsity parameter is a matter of trial and error. This mapping was successfully applied to ABE and does a good job of reconstructing the missing frequency components.

The results presented in the paper are based on objective evaluation and some primary subjective listening experiments. Listening tests in controlled environment should be performed to evaluate and quantify the performance of the system. We will be investigating these extensions in the near future. We also intend to explore and exploit the quasi-stationary nature of speech by modeling input subspace $\mathcal{P}$ with a HMM.

6. REFERENCES

[1] S. Voran, "Listener ratings of speech passbands," *IEEE workshop on Speech Coding*, pp. 81–82, 1997.

[2] W. Yasukawa, "Signal restoration of broadband speech using nonlinera processing," *Proc. European Signal Processing Conference*, , no. 176-178, 1996.

[3] Y. Yoshida and M. Abe, "An algorithm to reconstruct wideband speech from narrowband speech based on codebook mapping," *ICSLP*, pp. 1591–1594, 94.

[4] A. Chennoukh, S.and Gerritis, G. Miet, and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies," *ICASSP 01*, pp. 665 – 668, 2001.

[5] C. Avendano, H Hermansky, and E. Wan, "Beyond nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," *EUROSPEECH*, pp. 165–168, 1995.

[6] K. Kalgaonkar and M. A. Clements, "Vocal tract area based artificial bandwidth extension,," *IEEE MLSP Cancun, Mexico*, 2008.

[7] M. Hosoki, T. Nagai, and A. Kurematsu, "Speech signal bandwidth extension and noise removal using subband hmm," *ICASSP 02*, pp. 245–248, 2002.

[8] K. Kalgaonkar and M. Clements, "Vocal tract area based formant tracking using particle filters.," *ICASSP 08*, , no. 3405-3408, 2008.

[9] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[10] Matthew Brand, "Pattern discovery via entropy minimization," *Proc. Artificial Intelligence and Statistics*, 1998.

[11] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jerey, and D. E. Knuth, "On the lambert w function," *Advances in Computational Mathematics*, pp. 329–359, 1996.

[12] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using mmse estimation based on a hidden markove model," *ICASSP 03*, , no. 680-683, 2003.