Memory-Based Approximation of the Gaussian Mixture Model Framework for Bandwidth Extension of Narrowband Speech

Amr H. Nour-Eldin, Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montréal, Québec, Canada
amr.nour-eldin@mail.mcgill.ca, peter.kabal@mcgill.ca

Abstract

In this paper, we extend our previous work on exploiting speech temporal properties to improve Bandwidth Extension (BWE) of narrowband speech using Gaussian Mixture Models (GMMs). By quantifying temporal properties through information theoretic measures and using *delta* features, we have shown that narrowband memory significantly increases certainty about highband parameters. However, as delta features are non-invertible, they can not be directly used to reconstruct highband frequency content. In the work presented herein, we embed temporal properties indirectly into the GMM structure through a memory-dependent tree-based approach to extend representation of the narrow band. In particular, sequences of past frames are progressively used to grow the GMM in a tree-like fashion. This growth approach results in reliable estimates for the GMM parameters such that Maximum Likelihood estimation is no longer necessary, thus circumventing the complexity accompanying high-dimensionality GMM training.

Index Terms: Bandwidth extension, GMMs, speech memory.

1. Introduction

In traditional telephone networks, speech bandwidth is limited to the 0.3–3.4 kHz range. As a result, narrowband speech has sound quality inferior to its wideband counterpart and has reduced intelligibility especially for consonant sounds. Wideband speech reconstruction through Bandwidth Extension (BWE) attempts to regenerate the highband (3.4–8 kHz) frequency content lost during the filtering processes employed in traditional networks, thereby providing backward compatibility with existing networks

BWE is based on the assumption that narrowband speech correlates with the highband signal, and thus, given some a priori information about the nature of this correlation, the higher frequency speech content can be estimated. Considerable research has been dedicated to modelling this correlation, typically through either codebook mapping, or Gaussian Mixture Models (GMMs). While codebook mapping techniques discretize the acoustic space through Vector Quantization (VQ), GMMs provide a continuous approximation, and hence, outperform VQ-based methods. First used in [1] for the purpose of BWE, GMMs provide minimum mean square error (MMSE) estimates for highband spectral envelopes. Using linear predictive (LP) techniques, the statistically-estimated LP coefficients of said envelopes can, then, be combined with a highband residual error (excitation) signal in an LP synthesis filter to regenerate the missing highband signal. This signal is, in turn, added to the available narrowband signal to generate wideband speech.

In contrast, the correlation assumption between the narrow and highband spectral envelopes has itself received less attention. In [2], the certainty about the high band given the narrow band was quantified by determining the ratio of the Mutual Information (MI) between the two bands to the discrete entropy of the high band. The authors show that this ratio (representing the dependence between the two bands) is quite low. The relation of this ratio to BWE performance was further confirmed in [3] by deriving an upper bound on achievable BWE performance—represented by log-*Spectral Distortion* (LSD)—given a certain amount of MI and highband entropy. Despite the low dependence, BWE schemes have, for the most part, continued to use *memoryless mapping* between spectra of both bands. These schemes perform reasonably, however, not because they accurately predict the true high band, but rather by extending the narrow band such that the overall wideband signal sounds pleasant. Exceptions to the pervasiveness of memoryless mapping in BWE are based mainly on the implementation of highband spectrum envelope estimation using Hidden Markov Models (HMMs); e.g., [4]. Such HMM-based techniques are, however, marked by higher complexity and training data requirements, which increase with the number of HMM states. To mitigate the potential complexity and data insufficiency problems, first-order Markov models are assumed. This limits such HMM-based techniques to modelling the temporal dependencies between consecutive signal frames only, effectively restricting the ability of the model to capture only 20–40 ms of memory. It has been shown, however, that speech temporal information may extend up to 1000 ms [5], with energies of modulation spectra (spectra of the temporal envelopes of the signal) peaking around 4–5 Hz corresponding to 200–250 ms [6].

In Section 2, we provide the motivation of the current work presented herein by summarizing our previous work on transferring memory-based information theoretic gains into practical BWE performance. In Section 3, we extend the GMM formulation to take account of speech memory, and present a novel tree-based growth technique to construct extended GMMs. Finally, in Section 4, we present BWE results using the constructed memory-dependent extended GMM.

2. Review of previous work and motivation

In [7], we extended the work of [2] by quantifying the role of speech memory in increasing certainty about the high band. In particular, we transformed the *static* spectral envelope parametrization into a *dynamic* one by making use of delta features to represent memory at varying lengths around static frames. Delta features can be applied to any form of parametrization, and provide two advantages over first-order Markov chains. First, they capture temporal dynamics around

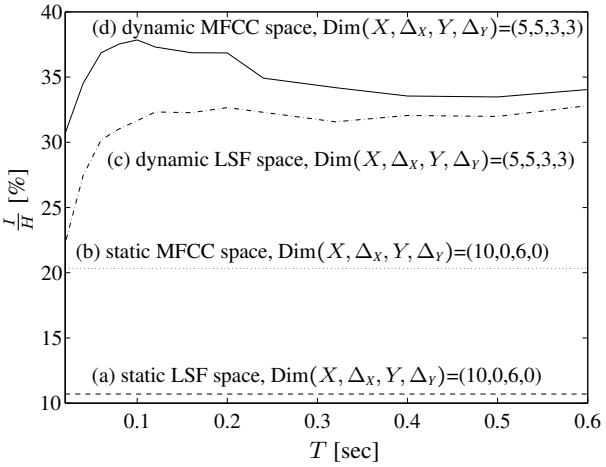

Figure 1: Highband certainty, $\frac{I}{H}$, versus span of memory, T, for static and dynamic (static+delta) MFCC/LSF feature spaces with narrow and highband dimensionalities of 10 and 6, resp.

static frames beyond the immediately neighbouring frames, and secondly, being a many-to-one transformation, they condense temporal information into a single vector that can be used to extend or replace part of the corresponding static feature vectors. This latter property eliminates increases in GMM complexity, and hence, requires no additional computational resources or data amounts required for statistical training.

Highband certainty is defined as the ratio of MI, $I(\cdot;\cdot)$, to discrete highband entropy, $H(\cdot)$. Representing static vectors of the narrow and high bands by X and Y, respectively, with Δ_X and Δ_Y representing the corresponding delta coefficient vectors, the highband certainty obtained with static and dynamic front-ends can then be written as $\frac{I(X;Y)}{H(Y)}$ and $\frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)}$, respectively. Using GMMs and VQ of the highband feature vectors to estimate I and H, respectively, we obtained the highband certainty results illustrated in Fig. 1 for varying widths, T, of the time window used to calculate delta features (dynamic spaces are obtained by replacing half the static features used in the reference static spaces by delta ones). Figure 1 shows the considerable highband certainty gains achieved by memory inclusion, yet with no increase in front-end dimensionality. The gains peak for $60 \lesssim T \lesssim 260$ ms, which includes the $200 \lesssim T \lesssim 250$ ms syllabic range. These results, thus, agree with the modulation spectra findings of [6] which show speech information content to be highest at the syllabic rate of 4–5 Hz.

In [8], we constructed a BWE system that makes use of delta features to take advantage of the highband certainty gains shown above. Average LSD performance improvements of approximately 5–7% were shown. While these improvements seem modest compared to the considerable information theoretic gains of Fig. 1, it should be noted that LSD performance is as strongly dependent on the GMM-based gain estimates used to generate excitation, as it is dependent on the spectral envelope LPC estimates. The highband certainty gains measured above consider only spectral envelopes. As such, information theoretic gains should rather be viewed as upper bounds on LSD performance. Moreover, as delta features are obtained by non-causal FIR filtering of static features with zeroes on the unit circle (differentiator), they are not practically invertible as the inverse filter is marginally stable. Consequently, delta features can not be used for LPC reconstruction, translating into a decrease of static feature dimensionality when overall highband dimensionality is fixed (as in the dynamic systems of Fig. 1). The resulting loss in highband spectral information is, however, compensated by the larger MI gains emanating from the GMM's superior cross-band covariances (since delta features have higher cross-band correlation compared to static features).

These drawbacks represent the motivation to pursue memory exploitation through a different avenue. In particular, we seek a technique that preserves highband dimensionality, does not require increases in training data requirements, and further considers only causal memory for the benefit of real-time implementation. Such a technique should also provide flexibility in regards to the extent of higher-order memory modelled; the primary advantage of delta features and deficiency of first-order HMM-based methods.

3. Memory-based GMM formulation

3.1. GMM definition

GMMs model the joint density of two random variables; x and y, (narrow and highband representations, resp., in the context of BWE) as a mixture of M component densities; i.e.,

$$f_{\text{GMM}}(x,y) = \sum_{m=1}^{M} \alpha_m f_{\text{G}}(x,y|\lambda_m), \tag{1}$$

where M is the number of mixture components, α_m is the mth mixture weight (prior probability) and $f_{\text{G}}(\cdot)$ denotes the multivariate Gaussian distribution defined by the mean vector μ_m and covariance matrix C_m in $\lambda_m = \{\mu_m, C_m\}$. Typically, the GMM parameters; (α_m, λ_m), are estimated by Maximum Likelihood (ML) estimation iteratively using the popular Expectation-Maximization (EM) algorithm.

3.2. Joint density MMSE estimation

For a known x, the function $\hat{y}_{\text{mse}} = F(x)$ that minimizes the mean square error; $\varepsilon_{\text{mse}} = E\left[\|y - F(x)\|^2\right]$, is the Expectation (dropping the subscript in p_{G})

$$\begin{aligned} E[\,y\,|\,x\,] &= \int_{\mathcal{R}_y} y\, p(y|x)\, dy = \int_{\mathcal{R}_y} y\, \frac{p(y,x)}{p(x)}\, dy, \\ &= \int_{\mathcal{R}_y} y\, \frac{\sum_{i=1}^{M} \alpha_i\, p(y,x|\lambda_i)}{\sum_{j=1}^{M} \alpha_j\, p(x|\lambda_j)}\, dy, \\ &= \frac{\sum_{i=1}^{M} \alpha_i\, p(x|\lambda_i) \int_{\mathcal{R}_y} y\, p(y|x,\lambda_i)\, dy}{\sum_{j=1}^{M} \alpha_j\, p(x|\lambda_j)}, \\ &= \sum_{i=1}^{M} h_i(x)\, E[y|x,\lambda_i], \end{aligned} \tag{2}$$

and since $p(y|x,\lambda_i) \triangleq f_{\text{G}}(y|x,\lambda_i)$ is Gaussian by definition in Eq. (1), $\therefore E[y|x,\lambda_i] = \mu_i^y + C_i^{yx} C_i^{xx^{-1}} (x - \mu_i^x)$.

3.3. Extending the GMM

In the GMM definition and MMSE formulation detailed above, it is assumed that the narrow and highband representations are static; i.e., x and y can be written as x_t and y_t. By fixing the dimensionality of the unknown highband representation; y_t, the straightforward extension of the formulation to include first-order memory can be obtained by substituting the

static narrowband representation; x_t, by an augmented version: $[x_t, x_{t-\tau}]^T$, using the narrowband features at time $t-\tau$, and replacing the joint densities of $[y_t, x_t]^T$—defined by (I, α_i, λ_i) for $i = 1, \ldots, I$—by (K, α_k, λ_k) where $k = 1, \ldots, K$, $K \geq I$ and $(K, \alpha_k, \lambda_k) \perp\!\!\!\perp [1] (I, \alpha_i, \lambda_i)$. While theoretically simple, this extension, in practice, treats the extended GMM—(K, α_k, λ_k), modelling the joint density of $\mathbf{z}^T \triangleq [y_t, x_t, x_{t-\tau}]^T$—as a completely new GMM to be EM trained without making any use of the prior knowledge about $[y_t, x_t]^T$ readily available in (I, α_i, λ_i). Moreover, the increased narrowband dimensionality requires a proportional increase in training data[2].

Alternatively, we propose the following approximation to (K, α_k, λ_k) by exploiting the prior knowledge in the reliably ML-estimated (I, α_i, λ_i). Let (J, α_j, λ_j), where $j = 1, \ldots, J$, represent the marginal density of x_t trained independently; i.e., $(I, \alpha_i, \lambda_i) \perp\!\!\!\perp (J, \alpha_j, \lambda_j)$. Given the lower dimensionality of x_t compared to $[y_t, x_t]^T$, we are guaranteed that—for $J \leq I$ and same amount of training data—(J, α_j, λ_j) ML estimates are, at least, as reliable as those of (I, α_i, λ_i). Viewing the component Gaussians in (I, α_i, λ_i) as *parent* states at time t, and those in (J, α_j, λ_j) as potential *child* states at time $t-\tau$ (using the well-accepted assumption that speech is locally stationary), we can approximate (K, α_k, λ_k) by

$$(\hat{K}, \hat{\alpha}_k, \hat{\lambda}_k) = (I \cdot J, \alpha_{ij}, \lambda_{ij}) \qquad \forall \;\; i = 1, \ldots, I; \;\; j = 1, \ldots, J. \quad (3)$$

Rather than estimate $(\alpha_{ij}, \lambda_{ij})$ through EM, we make use of the available training data, as well as the information in (I, α_i, λ_i) and (J, α_j, λ_j), by partitioning training data into subsets Q_{ij}, with each subset consisting of the augmented data; $\mathbf{z}^T$, with maximum joint likelihood given the density pair $(\alpha_i, \alpha_j, \lambda_i, \lambda_j)$; i.e.,

$$Q_{ij} \triangleq \left\{ \mathbf{z} \in \mathcal{R}_{\mathbf{z}} : \begin{array}{c} \arg\max\limits_{\lambda_t \in \{\lambda_i\}} p(y_t, x_t | \lambda_t) = \lambda_i \;, \\ \arg\max\limits_{\lambda_{t-\tau} \in \{\lambda_j\}} p(x_{t-\tau} | \lambda_{t-\tau}) = \lambda_j \end{array} \right\}, \quad (4)$$

which we then use to generate estimates for $(\alpha_{ij}, \lambda_{ij})$ per

$$\hat{\alpha}_{ij} = \frac{N(Q_{ij})}{\sum\limits_{i,j} N(Q_{ij})} \;, \quad (5a)$$

$$\hat{\mu}_{ij} = \frac{1}{N(Q_{ij})} \sum_{\mathbf{z} \in Q_{ij}} \mathbf{z} \;, \quad (5b)$$

$$\hat{C}_{ij} = \frac{1}{N(Q_{ij})} \sum_{\mathbf{z} \in Q_{ij}} (\mathbf{z} - \mu_{ij})(\mathbf{z} - \mu_{ij})^T \;. \quad (5c)$$

Under the local stationarity assumption, we exploit higher-order memory by extending the partitioning step of Eq. (4) using (J, α_j, λ_j) at progressive time shifts of $t - l\tau$ where $l = 1, \ldots, L$; i.e., the augmented data subsets generalize to

$$Q_{i^t j^{t-\tau} \ldots j^{t-L\tau}} \triangleq \left\{ \mathbf{z} \in \mathcal{R}_{\mathbf{z}} : \begin{array}{c} \arg\max\limits_{\lambda_t \in \{\lambda_i\}} p(\; y_t, x_t | \lambda_t) = \lambda_{i^t} \;, \\ \arg\max\limits_{\lambda_{t-\tau} \in \{\lambda_j\}} p(\; x_{t-\tau} | \lambda_{t-\tau}) = \lambda_{j^{t-\tau}} \;, \\ \vdots \\ \arg\max\limits_{\lambda_{t-L\tau} \in \{\lambda_j\}} p(x_{t-L\tau} | \lambda_{t-L\tau}) = \lambda_{j^{t-L\tau}} \;, \end{array} \right\}. \quad (6)$$

[1] $\perp\!\!\!\perp$ denotes *independent of*.

[2] Empirically, a hundred data points are typically needed to obtain reliable estimates of each GMM parameter, see Eq. (6) in [7] for details.

Using $\left(\hat{\alpha}_{i^t j^{t-\tau} \ldots j^{t-L\tau}}, \hat{\lambda}_{i^t j^{t-\tau} \ldots j^{t-L\tau}}\right)$ obtained by substituting Eq. (6) into Eq. (5), the derivation of the MMSE estimate $\hat{y}_{t_{\text{mse}}} = F(x_t, x_{t-\tau}, \ldots, x_{t-L\tau})$ can be performed in a manner similar to that of Eq. (2). In particular, it can be shown that (for ease of notation, we simplify $[x_t, x_{t-\tau}, \ldots, x_{t-L\tau}]$ by $\mathbf{x}_{\tau,L}$, $i^t j^{t-\tau} \ldots j^{t-L\tau}$ by $i\mathbf{j}^{\tau,L}$, and dropping the hats on $\hat{\alpha}$ and $\hat{\lambda}$)

$$\hat{y}_{t_{\text{mse}}} = \sum_{i\mathbf{j}^{\tau,L}} h_{i\mathbf{j}^{\tau,L}}(\mathbf{x}_{\tau,L}) \cdot E\left[y_t | \mathbf{x}_{\tau,L}, \lambda_{i\mathbf{j}^{\tau,L}}\right], \quad (7)$$

where

$$h_{i\mathbf{j}^{\tau,L}}(\mathbf{x}_{\tau,L}) = \frac{\alpha_{i\mathbf{j}^{\tau,L}} \;\; p(\mathbf{x}_{\tau,L} | \lambda_{i\mathbf{j}^{\tau,L}})}{\sum\limits_{i\mathbf{j}^{\tau,L}} \alpha_{i\mathbf{j}^{\tau,L}} \;\; p(\mathbf{x}_{\tau,L} | \lambda_{i\mathbf{j}^{\tau,L}})}, \quad (8)$$

and

$$E\left[y_t | \mathbf{x}_{\tau,L}, \lambda_{i\mathbf{j}^{\tau,L}}\right] = \mu^y_{i\mathbf{j}^{\tau,L}} + \frac{C^{y\mathbf{x}_{\tau,L}}_{i\mathbf{j}^{\tau,L}}}{C^{\mathbf{x}_{\tau,L}\mathbf{x}_{\tau,L}}_{i\mathbf{j}^{\tau,L}}} (\mathbf{x}_{\tau,L} - \mu^{\mathbf{x}_{\tau,L}}_{i\mathbf{j}^{\tau,L}}). \quad (9)$$

An L-order GMM constructed progressively using Eqs. (6) and (5) is denoted by $(I \cdot J^L, \alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$

3.4. Addressing data insufficiency concerns by pruning

For an l-order GMM with $K_l = IJ^l$ densities and a training data set of fixed size, the occupancy rates $N(Q_{i\mathbf{j}^{\tau,l}})$ decrease exponentially as l increases[3]. As the reliability of the $(\alpha_{i\mathbf{j}^{\tau,l}}, \lambda_{i\mathbf{j}^{\tau,l}})$ estimates is strongly dependent on occupancy rates, we must ensure that $N(Q_{i\mathbf{j}^{\tau,l}}) \geq N_{\min} \;\; \forall \;\; i\mathbf{j}^{\tau,l}$. Accordingly, at each step l in the progressive GMM-tree construction, we prune the total number of densities; K_l by merging all *child* densities of the same immediate *parent* into a single *pass-through* density; i.e., $\forall \; i\mathbf{j}^{\tau,l}$: if $N(Q_{i\mathbf{j}^{\tau,l}}) < N_{\min}$, set

$$J_{t-l\tau} = 1, \quad (10a)$$

$$Q_{i\mathbf{j}^{\tau,l}} = Q_{i\mathbf{j}^{\tau,l-1}}, \quad (10b)$$

re-estimate Eqs. (5), then set

$$C^{y\mathbf{x}_{\tau,l}}_{i\mathbf{j}^{\tau,l}} = \begin{bmatrix} C^{y\mathbf{x}_{\tau,l-1}}_{i\mathbf{j}^{\tau,l-1}} \\ 0^{\text{Dim}(y_t) \times \text{Dim}(x_t)} \end{bmatrix}. \quad (10c)$$

Thus, Eqs. (10) ensure that the pruned pass-through density is identical to its parent in terms of its contribution to $\hat{y}_{t_{\text{mse}}}$; i.e., for the affected $i\mathbf{j}^{\tau,l}$: $\alpha_{i\mathbf{j}^{\tau,l}} \equiv \alpha_{i\mathbf{j}^{\tau,l-1}}$ and $\lambda_{i\mathbf{j}^{\tau,l}} \equiv \lambda_{i\mathbf{j}^{\tau,l-1}}$.

3.5. Reliability of the memory-based $(\alpha_{i\mathbf{j}^{\tau,l}}, \lambda_{i\mathbf{j}^{\tau,l}})$

Reliability of the $(\alpha_{i\mathbf{j}^{\tau,l}}, \lambda_{i\mathbf{j}^{\tau,l}})$ estimates was empirically measured by comparing LSD performance on test data[4] over the following pairs of GMMs;

(a) a set of memory-based GMMs; $\left\{(\alpha_{i\mathbf{j}^{\tau=1,L}}, \lambda_{i\mathbf{j}^{\tau=1,L}})\right\}$, with $I = 128$, $J = 2$, $L = 1, \ldots, 5$, and $N_{\min} = 100$;

(b) an independent set of GMMs; $\left\{(\alpha_k, \lambda_k)\right\}$, with dimensionalities and number of densities equal to those in (a), but k-means initialized and EM trained with 100 iterations.

For all five pairs, the EM-trained GMMs of set (b) outperformed the memory-based ones of set (a) by a relative difference of $d_{\text{LSD}} < 2\%$, confirming reliability of the memory-based approximations of Section 3.3.

[3] Suitable I values for static GMM-based BWE systems are $I = 64, 128$. Even with $J_{\min} = 2$, the potential for data insufficiency problems for $L \gtrsim 2$ is clear.

[4] See Section 4 for details on the BWE system and data used.

4. Implementation and results

4.1. MFCC-based BWE system description

We showed in [9] that high-quality highband speech reconstruction from MFCCs is feasible using a simple cepstral-domain interpolation scheme based on high-resolution inverse Discrete Cosine Transform (IDCT). Through this scheme, we were able to exploit the superior correlation properties of MFCCs—as shown by Fig. 1—to implement a static MFCC-based BWE system that outperforms conventional LP-based ones. This BWE system also exploits equalization to extend the bandwidth of narrowband speech up to 4 kHz which allows extraction of an enhanced excitation signal required for highband LP-synthesis as described in Section 1. In addition, an excitation gain ratio, g_r, is used to scale the synthesized highband components such that their energy is equal to that of the corresponding frequency band in the original wideband speech used for GMM training. Being a perceptual property, this gain improves the subjective quality of the extended speech. It is also statistically estimated from narrowband parameters.

Accordingly, we reuse this system by constructing two memory-based GMMs $(\alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$; where y is the highband MFCC representation in the first GMM, and $y = g_r$ in the second. Training is performed using 20 msec frames with 50% overlap extracted form the TIMIT database (~3.1 hrs of training data and ~16 mins of core test data used on this analysis).

4.2. Results and analysis

We evaluate BWE performance in the missing 4–8 kHz band by LSD (dB);

$$d^2_{\text{LSD}} = \frac{1}{\pi}\int_{\omega_l}^{\omega_h}\left(20\log_{10}\frac{g}{|Y(e^{j\omega})|} - 20\log_{10}\frac{\hat{g}}{|\hat{Y}(e^{j\omega})|}\right)^2 d\omega,$$

where ω_l and ω_h are the cutoff frequencies of the missing high band, g and $Y(e^{j\omega})$ are the highband gain and frequency spectrum of the original wideband signal, respectively, while $\hat{g}$ and $\hat{Y}(e^{j\omega})$ are those of the GMM-estimated reconstructed signal.

While LSD is widely used for evaluating spectral envelope degradation due to its tractability and historic value, it does not take into account the perceptual importance of some aspects of the LP speech spectrum representation (e.g., it weights bandwidth differences for formants and valleys equally). Accordingly, we extend our performance analysis using PESQ; Perceptual Evaluation of Speech Quality [10], which was developed to model subjective tests commonly used in telecommunications, particularly mean opinion scores (MOS) covering a scale of 1 (bad) to 5 (excellent).

Figure 2 shows LSD as well as PESQ results for two BWE implementations of the memory-based GMMs $(\alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$; represented by △ and □, both with $I = 128$, $N_{\min} = 500$, and $\tau = 4$ (i.e., wideband vectors are augmented using past narrowband information at 40 msec shifts), while $J_\triangle = 2$ and $J_\square = 4$. Based on the shown results, we conclude:

1. Our memory-based approach results in clear performance improvements, both objectively as demonstrated by LSD improvements, as well as perceptually as indicated by PESQ results. Maximum LSD improvements reach ~ 12–13%, while PESQ ones reach ~ 8–10%. These relative improvements exceed those achieved by memory inclusion through delta features as described in [8].
2. Both systems reach a steady-state level of improvement at ~ 120–200 msec, coinciding with the region of maximum highband certainty gains measured in Fig. 1.

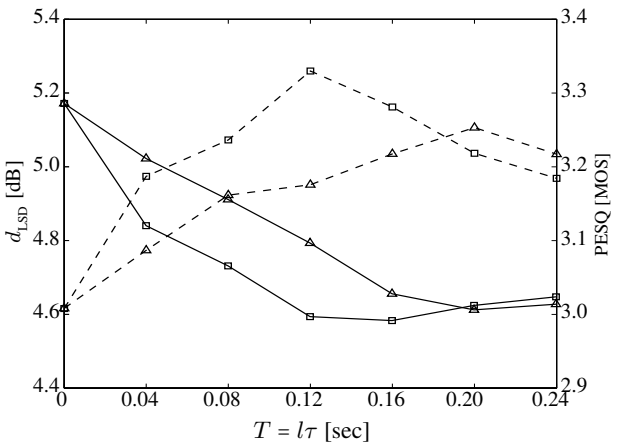

Figure 2: LSD and PESQ performance (solid and dashed lines, resp.) versus order of memory, $T = l\tau$, for the two BWE systems $(\alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$; △ and □, with $J_\triangle = 2$ and $J_\square = 4$.

3. As expected, system □ generally outperforms system △ for the same order of memory. This follows from the the fact that $J_\square > J_\triangle$ (with more degrees of freedom available for better modelling), and confirms the reliability of our tree-based GMM construction approach.
4. The importance of pruning to ensure reliability of the progressively growing $(\alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$ becomes quite clear by noting the variability of improvement rate with increasing $T = l\tau$ (and consequently, larger K). Particularly, the withdrawal of performance for higher-order memory indicates that the chosen pruning threshold; $N_{\min} = 500$ fails to compensate for the error in $(\alpha_{i\mathbf{j}^{\tau,L}}, \lambda_{i\mathbf{j}^{\tau,L}})$ estimates as K becomes overwhelmingly large.

5. References

[1] K.-Y. Park and H. S. Kim, "Narrowband to wideband conversion of speech using GMM based transformation", in *Proc. ICASSP*, Istanbul, pp. 1843–1846, 2000.

[2] M. Nilsson, H. Gustafsson, S. V. Andersen and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech", in *Proc. ICASSP*, Orlando, pp. 525–528, 2002.

[3] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals", in *Proc. ICASSP*, Orlando, pp. 237–240, 2002.

[4] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech", *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, 2003.

[5] H. Hermansky and S. Sharma, "TRAPS — Classifiers of temporal patterns", in *Proc. ICSLP*, Sydney, pp. 1003–1006, 1998.

[6] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech", in *Proc. ICASSP*, Munich, pp. 1647–1650, 1997.

[7] A. H. Nour-Eldin, T. Z. Shabestary and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands", in *Proc. ICASSP*, Toulouse, pp. III-53–56, 2006.

[8] A. H. Nour-Eldin and P. Kabal, "Combining frontend-based memory with MFCC features for Bandwidth Extension of narrowband speech", in *Proc. ICASSP*, Taipei, pp. 4001–4004, 2009.

[9] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech", in *Proc. InterSpeech*, Brisbane, pp. 53–56, 2008.

[10] ITU-T Recommendation P.862.2: *"Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs"*, November 2007.