Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (Toulouse, France), pp. III-53-III-56, May 2006

THE EFFECT OF MEMORY INCLUSION ON MUTUAL INFORMATION BETWEEN SPEECH FREQUENCY BANDS

Amr H. Nour-Eldin, Turaj Z. Shabestary and Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montréal, Québec, Canada

ABSTRACT

In this paper, we investigate the effect of temporal correlation on the dependence between the speech narrow and high frequency bands covering the 0.3–3.4 kHz and 3.7–8 kHz ranges, respectively. We follow the technique of using Gaussian mixture modelling of spectral envelopes represented by Mel-frequency cepstral coefficients. The correlation between the disjoint speech frequency bands is quantified through mutual information (MI) and its ratio to highband entropy. Speech exhibits considerable temporal correlation that is not explicitly accounted for by *static* parametrization of spectral envelopes. Including memory in speech parametrization (through *delta features*) incorporates such temporal information of speech in its modelling, and hence, MI gains are to be expected resulting in bandwidth extension with better performance. Results show that exploiting delta features can increase certainty about the highband (ratio of MI to highband entropy) by as much as 216% relatively, corresponding to an absolute increase of 12%.

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 it shows reduced intelligibility especially for consonant sounds. Wideband speech reconstruction through bandwidth extension (BWE) attempts to regenerate the low (20–300 Hz) and high band (3.4–8 kHz) signals lost during the filtering processes employed in traditional networks, thereby providing backward compatibility with existing networks.

In contrast to the abundance of published research on BWE techniques, few researchers have investigated the correlation assumption between the narrow and high band spectral envelopes. In [1], a rough lower bound on the MI between narrow and high frequency bands was derived. This initial attempt, however, did not present a meaningful conclusion in terms of BWE. This work was extended in [2] to quantify the remaining uncertainty of the high band given the narrow band by determining the ratio of the MI between the two bands to the entropy of the high-band. The authors show that this ratio (representing the dependence between the narrow and high bands) is quite low. Therefore, existing BWE schemes based on *memoryless* mapping between spectra of both bands perform reasonably, not because they accurately predict the true high band, but rather by extending the narrow band such that the overall wideband signal sounds pleasant. Accordingly, BWE methods should make use of some perceptual properties to ensure that the extended speech sounds pleasant. More recently, Jax and Vary [3] show that characteristics of the excitation of the input speech, such as gain or voicing, should also be included in the speech feature vectors.

More relevant to the work presented here is the implementation of high-band spectrum envelope estimation using Hidden Markov Models (HMMs) [4], with the advantage of embedding time correlation properties of speech into spectrum estimation. Although the authors objectively show the superior performance of their HMM technique compared to others, the gain of exploiting speech temporal information through HMMs has not been explicitly quantified. Worthy of note is also the work presented in [3] in which the effect of several parameterizations on BWE performance was investigated in terms of *class separability* as well as MI, although features representing temporal information were not considered.

In the work presented here, temporal information is explicitly accounted for through the so-called *delta features* widely used in speech recognition. These features are obtained through linearly weighted differences between neighbouring conventional *static* feature vectors. Similar to [2] and [5], MI and highband entropy are estimated using the numerical method of stochastic integration, where the marginal and joint distributions of the narrow and high band parameterizations are modelled by Gaussian mixture models (GMMs) for both static and extended (static+delta) acoustic spaces.

To verify the goodness of the extended space models, we model the marginal delta space distribution by a separate GMM. We also extract the portion corresponding to the delta subspace distribution from the extended space GMM, thus providing a second model for the marginal delta space which may be inaccurately trained due to ill-conditioning effects (since delta feature covariances are typically an order of magnitude lower than those of static ones). The log-likelihoods of these two models are estimated using the test data set and compared. Under several extended space conditions, the log-likelihood differences are found to be $\leq 0.12\%$, thus confirming the goodness of our extended space models.

Finally, we investigate the effect of the extent of embedded temporal information on MI estimates. This is implemented by performing the above experiment for varying widths of the window of static feature vectors involved in the estimation of the delta features.

2. MUTUAL INFORMATION ESTIMATION

Representing the narrow and high bands by the continuous (vector) variables X and Y, respectively, the mutual information can be written in terms of the joint and marginal *pdf*s as

$$I(X;Y) = \int_{\Omega_Y}\int_{\Omega_X} f_{XY}(x,y)\log_2\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)\mathrm{d}x\,\mathrm{d}y. \quad (1)$$

Following [2] and [5], we model the densities using the GMM $f_{GMM}(x,y) = \sum_{m=1}^{M}\alpha_m f_G(x,y|\theta_m)$, where M is the number of mixture components, α_m is the mth mixture weight and $f_G(\cdot)$ denotes the mutlivariate Gaussian distribution defined by the mean vector and covariance matrix in $\theta_m = \{\mu_m, C_m\}$.

Rewriting Eq. (1) as $I(X;Y) = \mathrm{E}\left[\log_2\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)\right]$ and replacing the expectation operator by the sample mean yields (for N samples) $I(X;Y) \approx \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{f_{XY}(x_n,y_n)}{f_X(x_n)f_Y(y_n)}\right)$. Thus, MI can be estimated (in bits) using numerical integration by substituting the *pdf*s for their GMM estimates; i.e.,

$$\hat{I}(X;Y) = \frac{1}{N}\sum_{n=1}^{N}\log_2\left(\frac{f_{GMM}(x_n,y_n)}{f_{GMM}(x_n)f_{GMM}(y_n)}\right). \quad (2)$$

Similarly, an estimate of the differential entropy of Y is obtained by

$$\hat{h}(Y) = -\frac{1}{N}\sum_{n=1}^{N}\log_2\left(f_{GMM}(y_n)\right). \quad (3)$$

Since $\hat{h}(Y)$ is susceptible to variable scaling, the discrete entropy $H(Y)$ provides a more consistent estimate of highband self-information. $H(Y)$ is approximated by

$$H(Y) \approx h(Y) - \log_2(\Delta^K), \quad (4)$$

where K is the dimension of the vector Y. This approximation is only valid if the quantization step-size Δ is small enough such that the *pdf* of Y can be considered flat in each quantization bin (high-rate assumption). Furthermore, since this approximation applies only to scalar quantizers, a certain distortion is introduced [2]. Using the MSE criterion, such distortion is given by $D = K\frac{\Delta^2}{12}$. As Y represents the highband spectral envelope, the square-root of D gives the *average spectral distortion* (SD). An average SD of 1 dB is considered the threshold of *spectral transparency* for the narrowband range [6]. Since level discrimination in hearing decreases by a small amount for highband frequencies, the 1 dB threshold—while being more stringent than necessary—can still be applied for highband entropy calculations [2]. Thus, $H(Y)$ can be estimated from Eq. (4) given: (a) $h(Y)$ obtained through Eq. (3), and (b) Δ satisfying an average SD of 1 dB.

3. SPEECH PARAMETRIZATION AND MODELLING

3.1. Parametrization

We parameterize the narrow and high bands by Mel-frequency cepstral coefficients (MFCCs). MFCCs were chosen since they can be directly related to *log-spectral-distortion* [7]; an objective speech quality measure widely used to assess the performance of spectral envelope quantizers. Furthermore, as MFCCs are calculated using a DCT, they have the desirable property of being decorrelated for different speech classes. MFCCs were shown to provide the highest class separability among most common spectral envelope parameters [3]. The advantage in terms of our work is that employing MFCCs results in GMMs with higher discriminative ability between different speech classes, which in turn results in more accurate modelling of the acoustic space, and hence, better MI estimates.

Rather than indirectly capture the temporal information of speech through state-transitions in HMMs or increasing the amount of overlap of speech frames, we include memory directly in the spectral envelope parametrization in the form of *delta* coefficients appended to the MFCC vectors. Delta coefficients are obtained from the *static* MFCC vectors by a first-order regression (time-derivative) implemented through linearly weighted differences between neighbouring MFCC vectors. Since immediately successive frames show only minor differences between their MFCC parameters, the trajectory of parameter variation with time is more accurately and easily identified as the time separation between the the involved static frames increases. Hence, the difference weights increase in proportion to the distance (in frames) between the two static vectors whose difference is being evaluated. We employ the HTK toolkit [8] to parameterize narrow and high band speech using the following formula to calculate the delta coefficients;

$$\delta_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2}, \quad (5)$$

where δ_t is a delta coefficient vector at time t computed in terms of the corresponding static cepstral coefficient vectors $c_{t-\theta}$ to $c_{t+\theta}$, and Θ specifies the number of neighbouring static frames to consider.

The TIMIT speech corpus [9] (containing speech sampled at 16 kHz) supplied the training and test data sets. Fifteen Mel-warped triangular filters acting on band-pass (0.3–3.4 kHz) filtered speech files were used to obtain narrowband MFCCs, while 5 Mel-warped triangular filters acting on high-pass (3.7–8 kHz) filtered speech files provided highband MFCCs. The 3.4–3.7 kHz range was discarded to avoid any dependencies between the narrow and high bands resulting from the filtering transition bands [2]. Delta coefficients were then calculated from the resulting MFCCs by Eq. (5) and appended to the static MFCC vectors for the extended space case. As the ratio of highband to narrowband energy represents an important measure of dependence between the narrow and high bands, the 0th cepstral coefficient c_0 (and its delta coefficient) was also appended to feature vectors of each band in the static (and extended space) case(s).

3.2. GMM modelling

To reduce computational requirements involved in GMM training, the TIMIT corpus was not divided into phoneme classes as is the case in [2], thus permitting the use of a single GMM to model all phonemes in a manner similar to [5]. This is justified by the arguments stated above regarding the class separability and discriminative ability of MFCCs as spectral envelope parameters. However, the drawback is that full covariance matrices must be used accompanied by an increase in the number of mixtures in order to ensure sufficient modelling of the approximately 40 English phonemes. Typically, a hundred data points are needed to obtain reliable estimates of each GMM parameter [2]. Consequently, given a fixed amount of N available samples, there is a tradeoff between the number of mixtures M to be used and the dimensionality d of the acoustic space being modelled, given by

$$M = \left\lfloor \frac{N}{100(1 + d + \frac{d(d+1)}{2})} \right\rfloor. \quad (6)$$

To increase the amount of available data, 20 msec frames with 50% overlap were extracted form the 3696 training and 1344 test speech files available in the TIMIT database, resulting in 1,126,746 training and 411,620 test frames. For a maximum dimensionality of 16 for the joint GMM representing the wide band (numerator in Eq. (2)), Eq. (6) yields 73 mixtures. Using the stochastic relation $L = \frac{1}{N}\sum_{n=1}^{N}\log_2\left(f_{GMM}(x_n,y_n)\right)$ for the GMM log-likelihood, we found empirically that increasing the number of mixtures to 256 results in a mere 0.2% increase in the joint GMM log-likelihood when using the test data set. Accordingly, as a compromise between increased computation and modelling accuracy, we chose the number of mixtures (107) corresponding to the midpoint of increase in log-likelihood as the M to be used in all GMMs in Eqs. (2) and (3).

4. EFFECT OF MEMORY INCLUSION

Let X (and Y) represent the static MFCC vectors, including c_0, of the narrow (and high) band(s), with Δ_X (and Δ_Y) representing the delta coefficient vectors, including δ_{c_0}, of the narrow (and high) band(s). Then, dropping the hat sign on $\hat{I}(\cdot,\cdot)$, we estimate the increase in MI; Δ_I, between the narrow and high bands as a result of memory inclusion in the following two scenarios:

1. Adding memory to narrowband modelling *only*, i.e., extending the narrowband acoustic space, yielding

$$\Delta_I = I(X, \Delta_X; Y) - I(X; Y), \tag{7}$$

2. Adding memory to *both* narrow and high band modelling, i.e., extending both narrow and high band spaces, yielding

$$\Delta_I = I(X, \Delta_X; Y, \Delta_Y) - I(X; Y). \tag{8}$$

For the first scenario, the relations between the information content of the X, Y and Δ_X acoustic spaces can be easily visualized through the Venn diagram of Fig. 1. Using Fig. 1, Δ_I of Eq. (7) can be written as $\Delta_I \equiv (\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_4) - (\mathcal{R}_1 \cup \mathcal{R}_2) = \mathcal{R}_4$ representing the additional gain in MI between the narrow and high bands as a result of exploiting narrow band temporal information. A similar illustration for Scenario 2 is, however, more complex.

Rewriting Eq. (7) in terms of the component GMMs as in Eq. (2) yields (dropping the subscript in $f_{GMM}(\cdot)$);

$$\Delta_I = \frac{1}{N}\sum_{n=1}^{N} \log_2 \left(\frac{f(x_n, \delta_{x_n}, y_n) f(x_n)}{f(x_n, \delta_{x_n}) f(x_n, y_n)} \right). \tag{9}$$

By a similar substitution, Δ_I from Eq. (8) can be estimated by

$$\Delta_I = \frac{1}{N}\sum_{n=1}^{N} \log_2 \left(\frac{f(x_n, \delta_{x_n}, y_n, \delta_{y_n}) f(x_n) f(y_n)}{f(x_n, \delta_{x_n}) f(y_n, \delta_{y_n}) f(x_n, y_n)} \right). \tag{10}$$

The increase in *certainty* about the high band due to memory inclusion can then be quantified by $\beta_1 \triangleq \frac{\Delta_I}{H(Y)}$ for Scenario 1, and $\beta_2 \triangleq \frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)} - \frac{I(X;Y)}{H(Y)}$ for Scenario 2. Hence, there are a total of 5 GMMs to train in Scenario 1 and 6 for Scenario 2 in order to obtain β_1 and β_2. As described in Section 3.2, although the same number of mixtures is used for all GMMs (which differ in dimensionality), this number (107) was determined based on the log-likelihood of the GMM with maximum possible dimensionality of 16. This ensures that training samples are sufficient to accurately estimate GMMs of dimensionality ≤ 16. Hence, since delta feature covariances are typically an order of magnitude lower than those of static ones, the only potential source of inaccurate GMM modelling is that of ill-conditioned covariance matrices of spaces extended with the Δ_X subspace (in Scenario 1) or with the Δ_X and Δ_Y subspaces (in Scenario 2). In other words, the delta subspaces may not be as accurately modelled as the static subspaces, which would result in lower estimates of Δ_I.

To verify the goodness of the extended space GMMs in Eq. (9) ($f(x_n, \delta_{x_n}, y_n)$ and $f(x_n, \delta_{x_n})$), we model the marginal delta space distribution; $f(\delta_{x_n})$, by a separate GMM, thus avoiding potential ill-conditioning effects on delta subspace GMM parameter estimation. In addition, we extract the portion corresponding to the delta subspace from the extended space GMM, providing a second model for the marginal delta space which may be inaccurately trained due to ill-conditioning effects. The log-likelihoods of these two models are estimated using the test data set and compared. In most of our experiments described below in Section 5, we use 5 static MFCCs for the narrow band and 3 for the high band, resulting in a 5-dim. GMM for $f(\delta_{x_n})$, a 10-dim. GMM for $f(x_n, \delta_{x_n})$, and a 13-dim. one for $f(x_n, \delta_{x_n}, y_n)$. The estimated likelihood discrepancy is a negligible 0.08% between the $f(\delta_{x_n})$ and $f(x_n, \delta_{x_n})$ GMMs, and 0.12% between the $f(\delta_{x_n})$ and $f(x_n, \delta_{x_n}, y_n)$ GMMs, thus confirming that our extended space models are as good as static space ones.

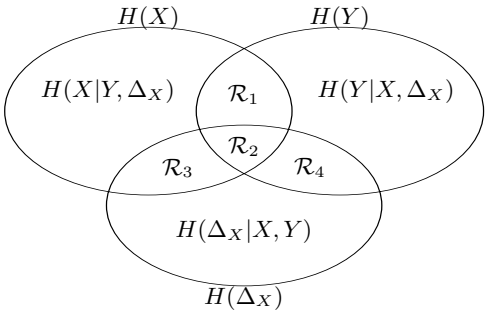

Fig. 1. Venn diagram representing the X, Y and Δ_X spaces.

Dim(X,Y)	$I(X;Y)$	$H(Y)$	$\frac{I(X;Y)}{H(Y)}$
$(10,6)$	1.52	27.42	5.55%
$(5,3)$	1.47	12.35	11.93%
$(10,3)$	1.49	12.35	12.04%

Table 1. Results showing the MI (in bits), highband entropy (in bits), and their ratio (in %) for the three reference static spaces.

5. SIMULATIONS AND RESULTS

We use a maximum dimensionality of 16 for joint narrow and high band GMM modelling (with maximum dimensionalities of 10 and 6 for the narrow and high bands, respectively) to reduce the computational requirements involved in GMM training. While this dimensionality is slightly lower than that used in [2], it should be noted that we employ full covariance matrices as opposed to diagonal ones as in [2]. The maximum narrowband dimensionality of 10 coincides with that of [5]. More importantly, it is the accurate estimation of MI differences (given by Eqs. (9) and (10)) that we seek rather than the exact effect of MFCC dimensionality on mutual information.

The dimensionalities of the extended spaces in our experiments are given by Dim$(X, \Delta_X, Y) = (5, 5, 3)$ and Dim$(X, \Delta_X, Y, \Delta_Y) = (5, 5, 3, 3)$ for Scenarios 1 and 2, respectively. We establish 3 reference *static* spaces for the estimation of $I(X;Y)$ against which to compare MI gains due to memory inclusion. These spaces are:

1. Dim$(X,Y) = (10, 6)$. The extended spaces of Scenario 2 are obtained by replacing the last 5 narrowband and 3 highband MFCCs by the delta coefficients of the first 5 narrowband and 3 highband MFCCs.
2. Dim$(X,Y) = (5, 3)$. The extended spaces are obtained by appending Δ_X coefficients for Scenario 1, and Δ_X and Δ_Y coefficients for Scenario 2.
3. Dim$(X,Y) = (10, 3)$. The extended space of Scenario 1 is obtained by replacing the last 5 narrowband MFCCs by the delta coefficients of the first 5.

Table 1 shows the information measure results obtained for these reference spaces. Comparing rows 2 and 3 shows that doubling the narrowband dimensionality for the same highband dimensionality only increases the $\frac{I(X;Y)}{H(Y)}$ ratio by 0.92%, thus demonstrating the

Scenario 1			
Θ	$I(X,\Delta_X;Y)$	$H(Y)$	$\frac{I(X,\Delta_X;Y)}{H(Y)}$
2	1.50	12.35	12.17%
≥ 4	1.47	12.35	11.93%
Scenario 2			
Θ	$I(X,\Delta_X;Y,\Delta_Y)$	$H(Y,\Delta_Y)$	$\frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)}$
2	2.76	21.29	12.96%
4	2.96	19.72	15.01%
8	2.87	17.61	16.33%
12	2.71	15.98	16.96%

Table 2. Results for the extended spaces of Scenarios 1 and 2.

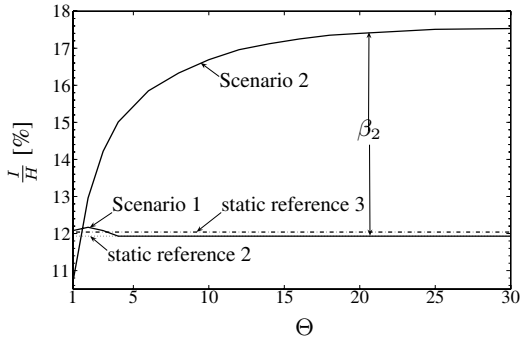

Fig. 2. Ratio of MI to highband entropy ($\frac{I}{H}$) versus Θ (span of frames used for calculation of delta coefficients) for the extended spaces of Scenarios 1 and 2. The $\frac{I}{H}$ ratio of static spaces 2 and 3 (rows 2 and 3 in Table 1, resp.) are also shown for reference.

low importance of narrowband dimensionality (for a specific highband dim.) on MI estimates. The considerably lower $\frac{I(X;Y)}{H(Y)}$ ratio for the first reference space compared to the other two is due to the additional information of the high band (resulting from the increased highband dim.), most of which is not shared with the narrow band.

The effect of memory inclusion on MI and highband certainty is represented in Table 2 for Scenarios 1 and 2. Fig. 2 provides a better illustration of the obtained results. By comparing the results pertaining to Scenario 1 with reference space 2 (i.e., where narrowband delta features are appended to static ones rather than replace them), it is clear that narrowband delta features add almost no new information about the static highband space. There is a modest maximum increase of 2% in highband certainity ($\beta_1 = 0.24\%$) for a window size of $\Theta = 2 \equiv 40$ msec (which adds a phoneme's transient onset and ending portions to its 20 msec steady-state portion). Hence, the modest increase results from additional information about the phonemes' identity. For $\Theta \geq 4$, however, there is no added information at all, i.e., $\beta_1 = 0$. In fact, comparing the same results to those of reference space 3 leads us to conclude that using additional narrowband static coefficients outperforms using delta coefficients.

In contrast, including memory in both narrow and high bands results in significant gains in both MI and highband certainty (β_2) as shown by the results pertaining to Scenario 2. Comparing these results with those of reference spaces 1 and 2 shows that replacing half of the static features by delta ones rather than appending them results in higher highband certainty. Fig. 2 clearly illustrates the strong effect of window size (Θ) on the $\frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)}$ ratio. The greatest gains are those due to short-term memory inclusion, i.e., for $\Theta \lesssim 8$ representing a memory of $t \lesssim 160$ msec. This duration corresponds roughly to triphones (phonemes with left and right contexts). In other words, the effect of memory inclusion is greatest when triphone-specific temporal information is employed to better identify individual phonemes (by exploiting inter-phoneme dependencies). Phonemes with mostly highband energy, e.g., fricatives, stand to have the most benefit of such short-term triphone memory inclusion. Since BWE schemes generally perform poorly when reconstructing such phonemes, the performance of such BWE schemes is expected to be considerably improved by triphone-specific memory inclusion. As shown in Fig. 2, the increase in highband certainty due to memory inclusion is slower for $8 \lesssim \Theta \lesssim 20$ ($160 \lesssim t \lesssim 400$ msec). This additional increase corresponds to intra-syllable dependencies. The gains in highband certainty level out at a $\frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)}$ ratio of approximately 17.53% for $\Theta \gtrsim 20$. This window size represents $t \gtrsim 400$ msec corresponding roughly to syllables. In other words, delta features fail to capture any additional information about inter-syllable dependencies. This is expected since such dependencies are determined by language-specific semantic construction rather than being a phonetic characteristic. Overall, memory inclusion per Scenario 2 results in a 215.9% maximum increase in highband certainty ($\beta_2 = 11.98\%$) when compared to reference space 1, and a 46.9% maximum increase ($\beta_2 = 5.6\%$) compared to reference space 2.

6. CONCLUSIONS

We have investigated the effect of temporal correlation on mutual information between speech narrow and high frequency bands. Temporal information is explicitly captured by delta coefficients of the MFCC-parameterized spectral envelopes. We have shown that such delta features efficiently incorporate relevant inter-phoneme as well as intra-syllable information, resulting in considerable gains in highband certainty. Accordingly, we conclude that memory inclusion provides an important source of potential BWE performance improvement at a negligible added computational cost.

7. REFERENCES

[1] M. Nilsson, S. V. Andersen and W. B. Kleijn, "On the mutual information between frequency bands in speech", in *Proc. ICASSP*, pp. 4085–4088, 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*, pp. 525–528, 2002.

[3] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals", in *Proc. ICASSP*, pp. 697–700, 2004.

[4] P. Jax and P. Vary, "Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model", in *Proc. ICASSP*, pp. 680–683, 2003.

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

[6] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. Speech, Audio Processing*, vol. 1, pp. 3–14, 1993.

[7] R. Hagen, "Spectral quantization of cepstral coefficients", in *Proc. ICASSP*, pp. 509–512, 1994.

[8] S. J. Young et al., *HTK Book–version 3.2*, Microsoft Corporation and Cambridge University Engineering Department, 2002.

[9] W. Fisher, et al., "The DARPA speech recognition research database: specification and status", in *Proc. DARPA Workshop on Speech Recognition*, pp. 93–99, 1986.