EXPLOITING EXPLICIT MEMORY INCLUSION FOR ARTIFICIAL BANDWIDTH EXTENSION

Pramod Bachhav, Massimiliano Todisco and Nicholas Evans

EURECOM, France

Emails: {bachhav, todisco, evans}@eurecom.fr

Introduction

- traditional telephony infrastructure is typically limited to a bandwidth of 0.3-3.4 kHz, referred as narrowband (NB)
- unvoiced phonemes exhibit significant information beyond NB
- wider bandwidths generally correspond to higher quality speech
- artificial bandwidth extension (ABE) methods estimate missing highband (HB) components at 3.4-8kHz
- use of dynamic information or *memory* to improve ABE performance is common and can be captured using back-end regression models or via front-end features
- memory inclusion via delta features for ABE has been investigated thoroughly [1,2] via information theoretic analysis
- a quantitative analysis of the benefit of *explicit* memory from neighbouring frames, without significant increases to complexity and latency is missing

Contributions

- assessment of *explicit* memory through information theoretic analysis
- *explicit* memory inclusion for ABE without affecting complexity of a standard regression model
- application of principal component analysis as a dimensionality reduction transform

Mutual information

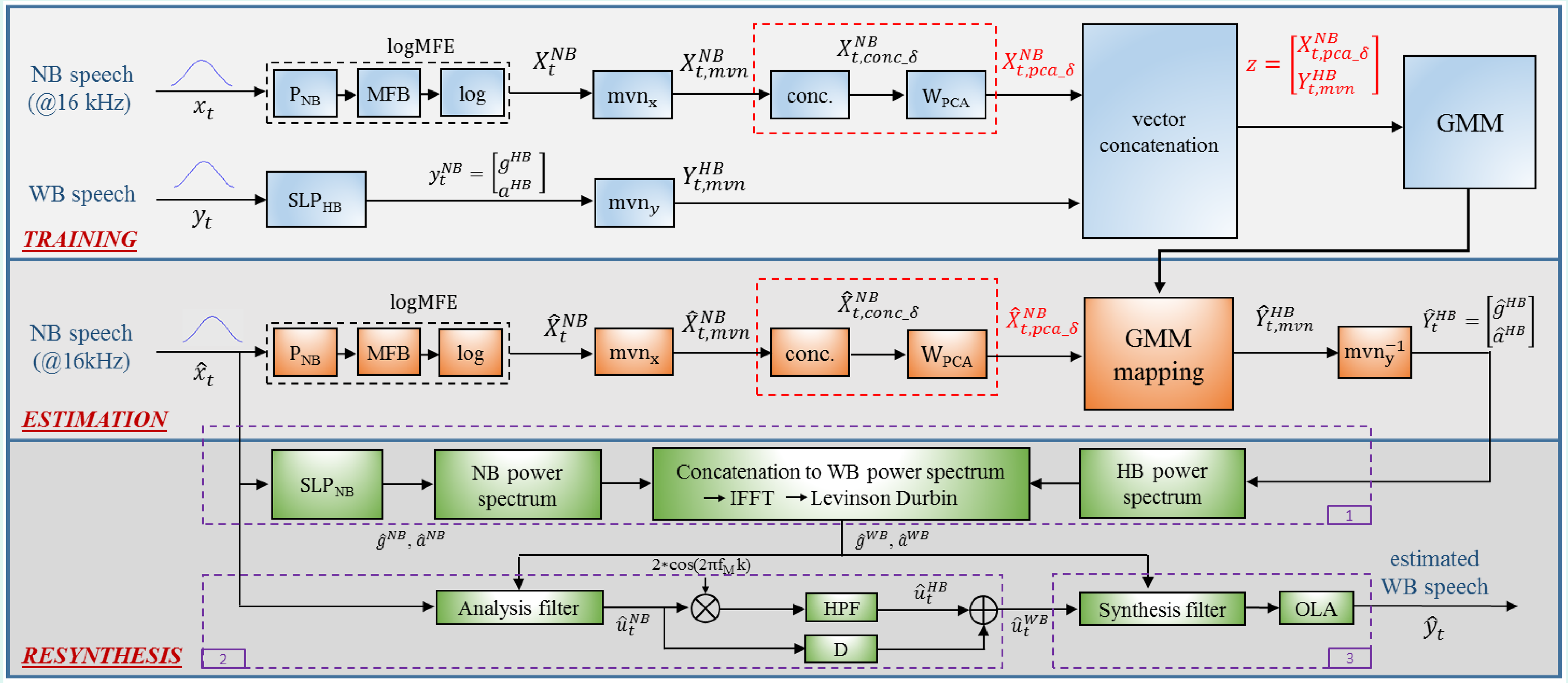
- correlation between NB and HB features is usually measured using mutual information (MI)
- the mutual information between two continuous random variables X and Y with joint probability density function (PDF) $f_{XY}(x,y)$ is defined according to:

$$I(X;Y) = \iint f_{XY}(x,y)\log_2\left(\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right)dxdy$$

- the integral can be written as an expectation approximated by the sample mean over K samples as follows:

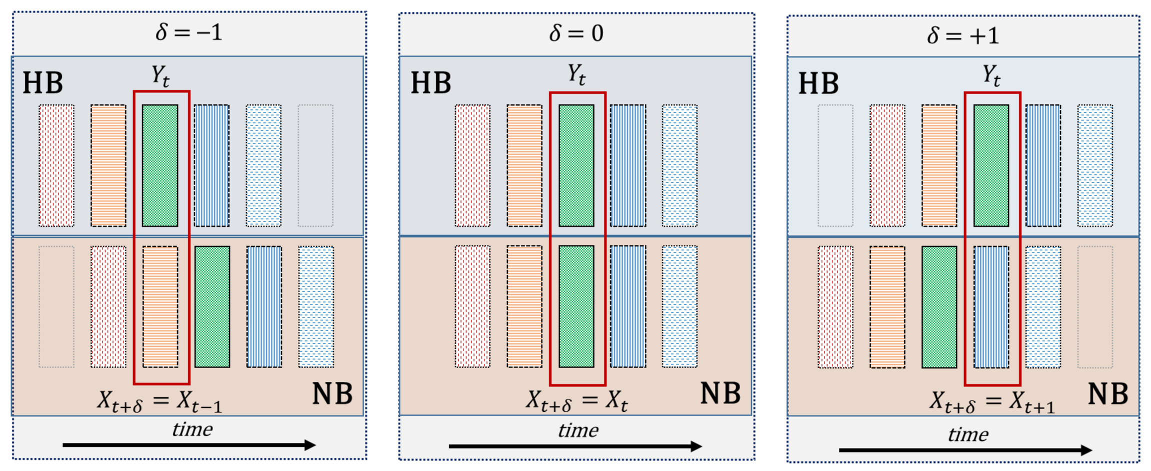
$$I(X;Y) \approx \frac{1}{K}\sum_{k=1}^{K}\log_2\left(\frac{f_{XY}(x_k,y_k)}{f_X(x_k)f_Y(y_k)}\right)$$

- the joint PDF $f_{XY}(x,y)$ is usually modelled using a Gaussian mixture model (GMM)
- A GMM of 128 components was used to estimate MI between NB and HB representations, using TIMIT database

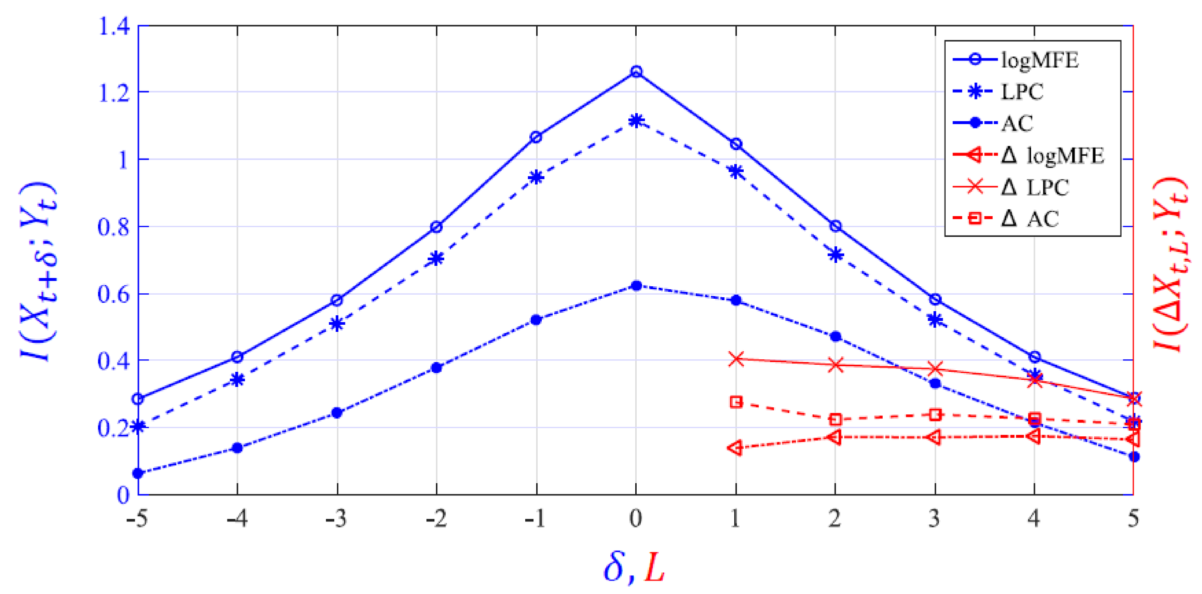


A block diagram of the ABE system with memory inclusion

Benefit of *memory* to ABE



An Illustration of MI estimation with contextual information from neighbouring frames. Vertical bars represent NB (bottom) and HB (top) feature vectors. Red boxes represent the pair of NB ($X_{t+\delta}, \delta = -1,0,1$) and HB ($Y = Y_t$) components used for MI calculations.



An illustration of the variation in MI between static HB features Y_t and static NB features (blue profiles) extracted from neighbouring frames $X_{t+\delta}$, and delta features $\Delta X_{t,L}$ (red profiles).

Experimental setup and results

- Database: TIMIT database divided into training (3696 utterances) and test (1344 utterances) sets.
- NB features: 10 log Mel filter (logMFE) coefficients; HB features: 10 linear prediction (LP) coefficients including LP gain
- Mapping: GMM regression [3] (using 128 components)
- Proposed ABE system with *memory* M_δ uses δ neighboring frames (NB features - $\hat{X}^{NB}_{t,pca\_\delta}$)
- Baseline B1 uses static NB features $\hat{X}^{NB}_t$
- Baseline B2 uses NB and HB features formed by appending 5 static features with corresponding 5 second order delta coefficients. (A variant of the approach presented in [2])

ABE method	$d_{\text{RMS-LSD}}$	d_{COSH}	MOS-LQO
B1	9.2 (1.2)	2.4 (0.7)	2.4
B2	10.1 (1.2)	3.6 (1.2)	2.2
M_1	8.2 (0.9)	2.2 (0.6)	2.8
M_2	**8.1** (0.9)	**2.1** (0.6)	**2.9**
M_3	8.2 (0.9)	2.2 (0.7)	2.8

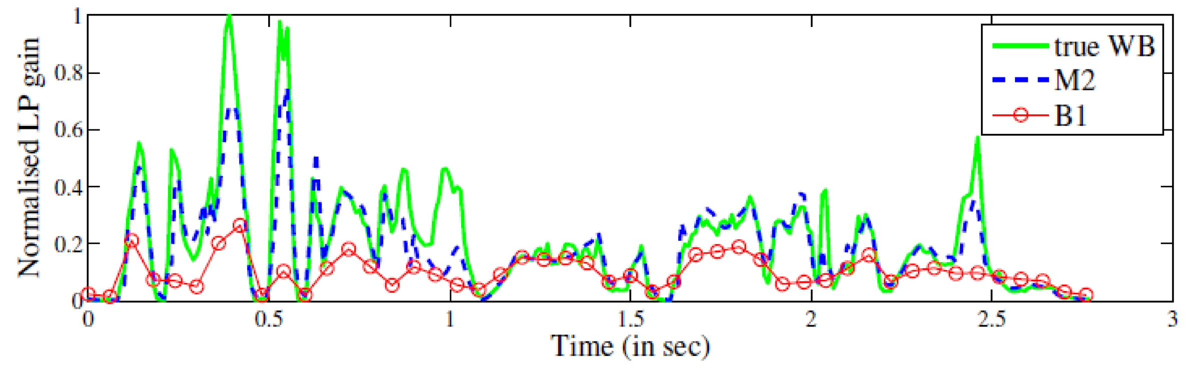
Objective assessment results. Lower values of $d_{\text{RMS-LSD}}$ and d_{COSH} indicate better performance whereas as higher MOS-LQO indicates better quality.

Comparison B → A	CMOS
$M_2 \rightarrow$ NB	0.69
$M_2 \rightarrow$ B1	0.51
$M_2 \rightarrow$ WB	-0.78

Subjective assessment results in terms of CMOS. Files used for the subjective evaluation are available at ***http://audio.eurecom.fr/content/media***

Comparison	logMFE
$I(X_t; Y_t)$	1.24
$I(X^{NB}_{t,pca\_2}; Y_t)$	1.34

Mutual information assessment results



A comparison of true WB LP gain $\hat{g}^{WB}_{true}$ to estimated WB LP gain $\hat{g}^{WB}$ for systems M_2 and B1

Conclusions and future work

- *explicit memory* inclusion for ABE is presented without significant impact on computational complexity
- use of PCA is as a dimensionality reduction transform
- potential of the *memory* is demonstrated through information theoretic analysis
- *memory* produces bandwidth-extended speech signals with better speech quality

Future Work

- investigation of dimensionality reduction techniques designed to preserve speech quality rather than feature variance

Selected References

[1] A. Nour-Eldin et al., "The effect of memory inclusion on mutual information between speech frequency bands", *Proc. of ICASSP*, 2006

[2] A. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narroband speech", in *Proc. of INTERSPEECH*, 2008

[3] K.-Y. Park and H. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," in *Proc. of ICASSP*, 2000

[4] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals", in *Proc. of ICASSP*, 2004

Presented by Pramod Bachhav at 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, Calgary, Canada.