EFFICIENT SUPER-WIDE BANDWIDTH EXTENSION USING LINEAR PREDICTION BASED ANALYSIS-SYNTHESIS

Pramod Bachhav, Massimiliano Todisco and Nicholas Evans

EURECOM, France

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

EURECOM
Sophia Antipolis

Introduction

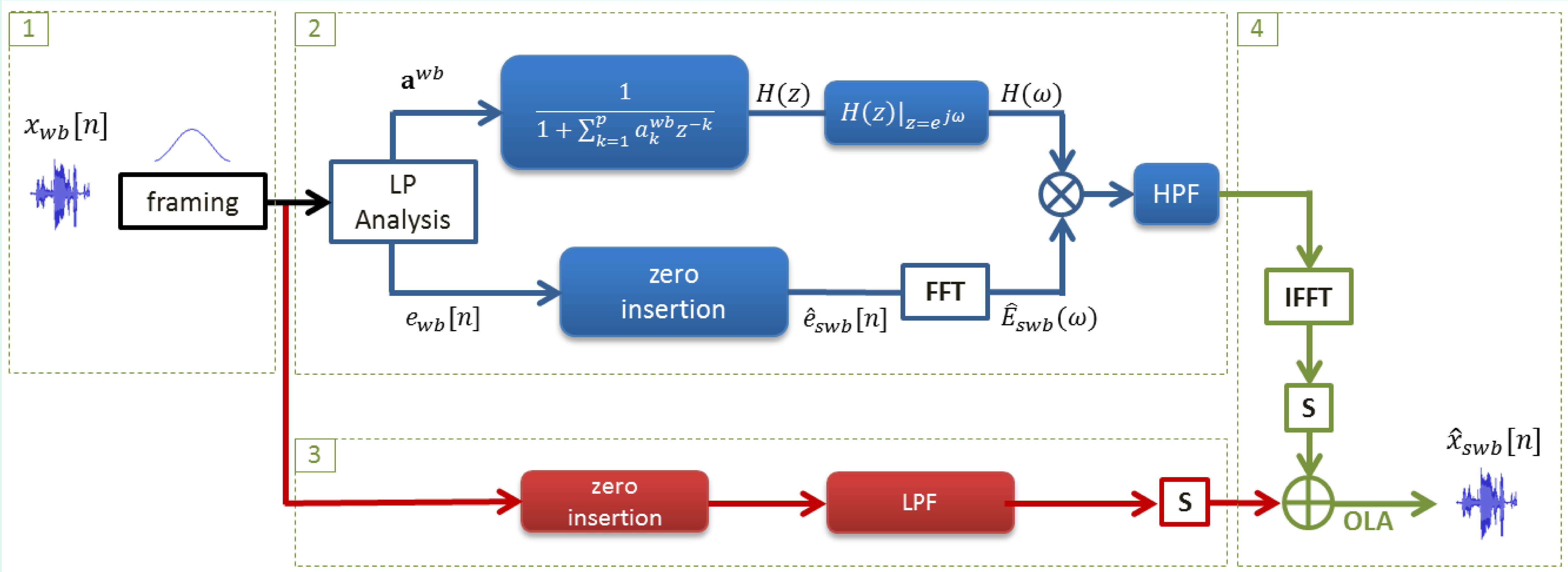
- traditional telephony infrastructure is typically limited to a bandwidth of 0.3-3.4 kHz, referred as narrowband (NB)
- wider bandwidths generally correspond to higher quality speech
- artificial bandwidth extension (ABE) methods estimate missing frequency components at 3.4-8kHz
- today's devices are capable of supporting wideband (WB) and super-wideband (SWB) communications at bandwidths 7kHz and 14kHz respectively
- the adaptive multi-rate WB (AMR-WB) and enhanced voice services (EVS) codecs are respective examples
- until all network services and devices move to super-wide bandwidth, SWB devices may often be restricted to NB or WB communications
- super-wide bandwidth extension (SWBE) approaches, therefore, are used to estimate missing high frequency (HF) components between 8-16kHz from available low frequency (LF) components between 0-8kHz

Past work

- ABE algorithms are usually classified as blind and non-blind
- non-blind algorithms perform ABE using auxiliary side HF information encoded with LF components
  - this extra information incurs an additional burden on bit rate
  - e.g. EVS codec (SWB mode), extended AMR-WB (AMR-WB+) codec, high efficiency advanced audio codec (HE-AAC)
- in contrast, blind algorithms use only the LF information
- most existing SWBE algorithms use statistical estimation techniques to predict the missing HF information
- this extra estimation step augments complexity and introduces latency

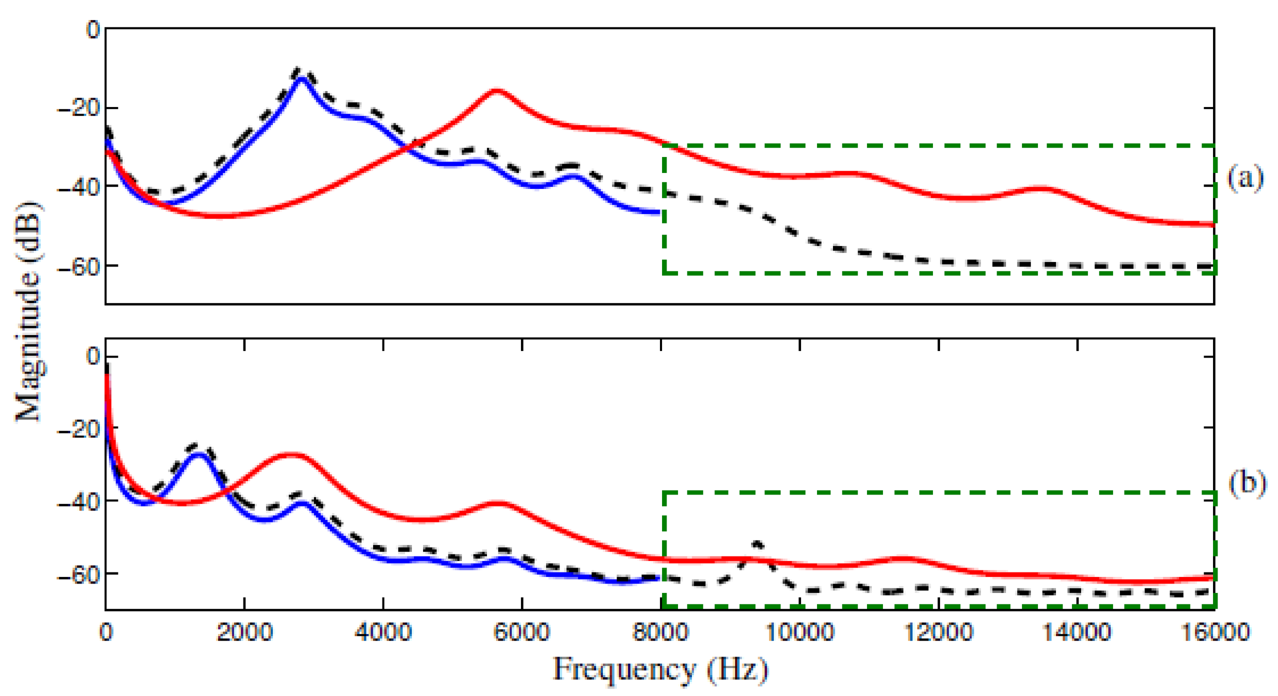
Contributions

- an efficient approach to SWBE based on linear prediction (LP) analysis synthesis
- the missing HF components are extracted from the WB-LP spectral envelope without any statistical estimation
- SWBE is performed without increasing complexity or latency
- performance is compared to a state-of-the-art EVS codec



A block diagram of the proposed approach to SWBE

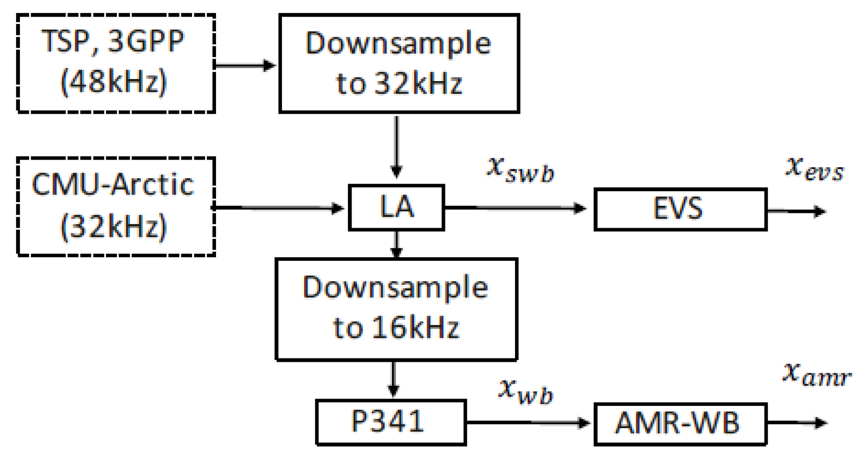
Spectral envelope analysis



A comparison of spectral envelopes for an arbitrary speech frame. Profiles shown for true WB speech (blue), true SWB speech (dashed black). Red profiles show the stretched copy of WB envelope, equivalent to effective frequency response for extended excitation. Plots shown for distinct frames of (a) unvoiced and (b) voiced speech

Experimental setup

- Databases: **CMU Arctic** database (1132 utterances at 32kHz, 3 speakers), **TSP speech** database (1378 utterances, 12 male and 12 female speakers), 6 English utterances chosen from **3GPP** database
- Baseline: efficient high-frequency bandwidth extension (EHBE) algorithm [1] implemented in time domain without framing
- baseline does not need any statistical estimation and performs comparable to recent approach presented in [2]
- input to both the proposed approach and the baseline are the WB signals processed with AMR-WB codec at 12.65 kbps
- extended signals using proposed approach are also compared to SWB signals processed with the EVS codec at 13.2kbps

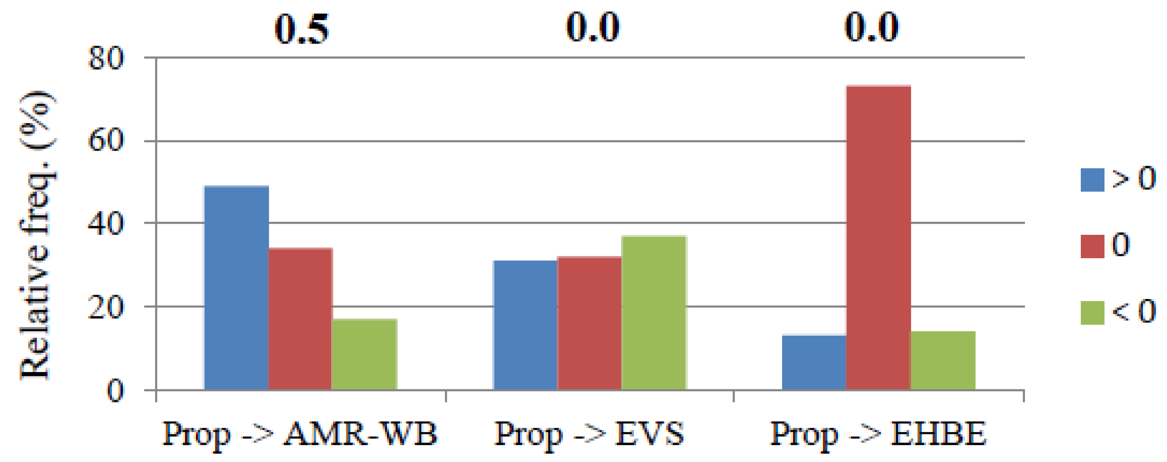


Protocol used for data pre-processing. LA = level alignment to -26 dBov.

Experimental results

	Proposed	EHBE	EVS
CMU Arctic	10.13 (1.68)	11.74 (2.03)	5.00 (0.48)
3GPP	11.06 (1.90)	13.56 (2.30)	4.87 (0.39)
TSP speech	9.29 (0.84)	10.20 (1.04)	4.74 (0.51)
Average	9.92 (1.56)	11.36 (1.96)	4.94 (0.50)

RMS-LSD results in dB (standard deviation).



Subjective test results in terms of CMOS for bandwidth extended speech generated with the proposed (Prop) algorithm (A) versus either AMR-WB, EVS and EHBE processed speech (B). Each bar indicates the relative frequency that (blue bars) A was preferred to B (score>0), that (green bars) quality was indistinguishable (score=0), or that (red bars) B was preferred to A (score< 0). Scores illustrated to the top are average subjective scores.

Discussion

- despite improvements in objective results, preference for the proposed approach is slightly lower than the EHBE baseline
- this is possibly because of implementation differences. Time domain processing used for the baseline without framing leads to less artefacts
- compared to RMS-LSD performance gap, preference for the EVS processed speech signals in subjective tests is marginal
- reduced level discrimination at higher frequencies [4] maybe the possible reason

Conclusions and future work

- a simple yet effective SWBE approach is presented
- no need for statistical estimation
- codec neutral
- could be more efficient than the baseline, if used with a codec employing some form of linear prediction (e.g. AMR-WB codec)
- future work: thorough investigation and comparison of complexity and latency for suitable real time implementations

Selected References

[1] E. Larsen et. al, "Efficient high frequency bandwidth extension of music and speech," in *112th Audio Engineering Society Convention*, 2002

[2] C.-C. Bao et.al, "A blind bandwidth extension method for audio signals based on phase space reconstruction," *EURASIP Journal on Audio, Speech, and Music Processing*, 2014

[3] "Codec for Enhanced Voice Services; Detailed algorithmic description (3GPP TS 26.445 ver. 13.4.0 rel. 13)," 2016

[4] M. Florentine et al., "Level discrimination as a function of level for tones from 0.25 to 16 kHz," *Journal of the Acoustical Society of America*, 1987.

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