Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech

Emma Jokinen[1,2], *Ulpu Remes*[1], *Marko Takanen*[1],
Kalle Palomäki[1], *Mikko Kurimo*[1], *Paavo Alku*[1]

[1]Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
[2]International Audio Laboratories Erlangen, Friedrich-Alexander Universität (FAU), Germany
emma.jokinen@aalto.fi

Abstract

In mobile communications, post-processing methods are used to improve the intelligibility of speech in adverse background noise conditions. In this study, post-processing based on modelling the Lombard effect is investigated. The study focuses on comparing different spectral envelope estimation methods together with Gaussian mixture modelling in order to change the spectral tilt of speech in a post-processing algorithm. Six spectral envelope estimation methods are compared using objective distortion measures as well as subjective word-error rate and quality tests in different near-end noise conditions. Results show that one of the envelope estimation methods, stabilised weighted linear prediction, yielded statistically significant improvement in intelligibility over unprocessed speech.
Index Terms: Intelligibility enhancement, telephone speech, Gaussian mixture model, spectral envelope estimation

1. Introduction

Post-processing is used in mobile communications to enhance the quality and intelligibility of speech that has been degraded by quantization or acoustical background noise. The background noise can be on the sending or receiving side of the connection, referred to as the far-end and the near-end noise scenario, respectively. This study focuses on the near-end noise scenario, where the decoded speech is assumed to contain only quantization distortion. The goal of post-processing is to make the speech stand out better from the background noise by enhancing its acoustic cues.

Several intelligibility enhancement methods have been proposed in previous studies. They are based, for example, on optimizing objective measures, such as the speech intelligibility index (SII) or the glimpse proportion (GP) [1–4], or re-allocating speech energy with simple high-pass filtering [5–7]. While these post-processing techniques have been shown to improve intelligibility, they are not capable of modelling the spectral changes that occur in natural speech, for instance, when vocal effort is increased in order to enhance loudness. An example of this is the Lombard effect which is observed when talkers modify their speaking style in an effort to increase the intelligibility of their speech in the presence of background noise [8]. The increased intelligibility in Lombard speech is a combination of several factors, such as flattening of the spectral tilt, slower speaking rate, and increased vocal intensity [9, 10]. The Lombard effect has been used in both post-processing [11–15] as well as in speech synthesis [16,17]. However, advanced data-driven methods have not been used to model the Lombard effect in post-processing algorithms that are to be used in speech transmission technology. This application area, unlike speech synthesis, sets stringent constraints on the algorithmic delay and computational load of the processing.

In this study, a post-processing method based on Gaussian mixture models (GMMs) is proposed to enhance speech intelligibility. The method adjusts the spectral tilt of the received narrowband speech frame using statistical dependencies of normal and Lombard speech. GMMs were selected for the mapping because of their flexibility and the ease of interpreting them. Additionally, they have been succesfully used for similar purposes in voice conversion [18]. In order to find the best parametric model for the spectral tilt, six previously developed methods to parametrize the spectral tilt in normal and Lombard speech are compared. The envelope estimates, their mapping, and performance in post-processing were evaluated objectively and in subjective intelligibility and quality tests in different noise conditions. The study constitutes a first step towards developing a post-processing method capable of modelling the natural Lombard effect with speech transmission in mobile devices as the target application.

2. Spectral tilt estimation methods

For estimating the spectral tilt, six different methods were evaluated. The main criterion in selecting these six candidates was that the methods selected could be utilized, not only in modelling the spectral tilt, but also in compensating for it as a part of a post-processing method.

Dumbing filter (DMF): the DMF method has been used for spectral tilt correction in voice conversion [19]. It is of the form $1/(1 - gz^{-1})2$, where coefficient g is obtained by computing the real root of polynomial $g^3 + 3c_1g^2 + (2 + c_2)g + c_1$. Here c_i is the ith autocorrelation coefficient.

Stabilised weighted linear prediction (SWLP): SWLP [20] is an all-pole modelling technique similar to weighted linear prediction (WLP) [21] in which the square of the residual is temporally weighted. Differently from WLP, stability of the SWLP all-pole filter is guaranteed. By using the short-time energy (STE) as the weighting function and by selecting the length of the STE window appropriately, SWLP produces smooth spectral envelopes even with higher-order models. In this work, the STE window length was set to 1 and the model order to 10.

Two-stage LP (2LP): the 2LP method was used in [13] to estimate and compensate for the spectral tilt of voiced speech. The approach first computes a 20th-order LP analysis to the input frame and this is then used as an input to a 6th-order LP analysis. To remove any remaining effects of formants, the obtained impulse response is windowed with an exponential window.

Table 1: *Average errors (in dB) of the spectral tilt estimation methods obtained by computing the log-spectral distortion between the long-term average spectra (LTAS) of the parametrizations and the LTAS of normal (NOR) and Lombard (LOM) speech for male (M) and female (F) speakers.*

		DMF	SWLP	2LP	2SLP	OCT	TSF
NOR	M	2.97	4.47	6.95	7.56	3.41	9.45
	F	3.34	3.94	6.42	6.97	4.11	8.56
LOM	M	3.36	3.36	4.46	4.44	4.83	5.50
	F	3.99	2.99	3.57	3.42	4.68	4.14

Two-stage selective LP (2SLP): 2SLP uses similar two-stage LP analysis as the 2LP method with the exception that the first LP analysis is computed with selective linear prediction [22]. Instead of fitting the LP model to the entire frequency range from 0 Hz to 4 kHz, the fit is computed only in the 300 Hz-4 kHz range by using linear mapping of the spectrum. This is motivated by the filtering that is done at the input of the mobile device, which strongly suppresses frequencies below 300 Hz. Without the linear mapping, large attenuation at the lowest frequencies might distort all-pole modelling of the spectral tilt.
1/3-octave band energy fit (OCT): the OCT method is based on [9] where spectral tilts of normal and Lombard speech were parametrized by a regression fit to spectral energies at 1/3-octave bands. In the present study, however, a 4th-order all-pole filter is computed from 1/3-octave band energies and only 15 bands ranging to 4 kHz are used. Based on the sub-band energies, E_i, a magnitude spectrum is constructed where each component in the ith sub-band is set to $\sqrt{E_i}/N_i$, where N_i is the number of components in the ith sub-band. Autocorrelation is computed from the magnitude spectrum and used to obtain a LP fit with the Levinson-Durbin recursion.
Telephone sub-band magnitude fit (TSF): the TSF method is derived from a spectral feature used for artificial bandwidth extension [23]. Originally, the narrowband speech spectrum is divided into four sub-bands (0.3-1.0 kHz, 1.0-1.7 kHz, 1.7-2.5 kHz, and 2.5-3.4 kHz) and the average spectral magnitude is computed in each sub-band. In this study, a fifth sub-band was added (3.4-4 kHz) and a 4th-order LP fit is computed utilizing the average spectral magnitudes as in the OCT method.

To simplify the comparison and processing, all spectral tilt estimation methods were designed to produce a digital filter. In the case of 2LP and 2SLP, the tilt is modelled with an all-zero filter, whereas in the case of the other methods (DMF, SWLP, OCT, TSF), an all-pole filter is used.

2.1. Objective evaluation

Performance of the spectral tilt estimation methods was evaluated by comparing the long-term average spectra (LTAS) of the all-pole models. LTAS were computed for each method by using both normal and Lombard speech from the GMM training database which is described in Section 3. LTAS computed from normal and Lombard speech were compared to those extracted from the spectral tilt estimation methods using the log-spectral distortion [24]. The scores obtained were averaged to get the final error measures shown in Table 1. In Fig.1, LTAS from a male speaker and the LTAS of the tilt estimates are visualized for both normal and Lombard speech. Based on the objective LTAS-based error measures, 2LP and 2SLP perform similarly and therefore, 2SLP was removed from further evaluations.

3. GMM training

Two previously recorded databases with normal and Lombard speech in Finnish were used as training and development data for the GMMs. Both of them contain parallel recordings, i.e., each sentence was produced by each speaker using normal and Lombard speaking styles. The training database contains approximately 360 2-3 seconds long sentences from 6 speakers (3 males). The development database contains short recordings of parallel normal and Lombard speech from 18 speakers (9 males) [25]. The Lombard material was recorded by feeding noise to the speaker's ears through headphones. Only the voiced speech was used in the present study.

All data were preprocessed to correspond to narrowband telephone speech by filtering with the MSIN filter [26] at 16 kHz. The MSIN filter is a high-pass filter designed to simulate mobile station input characteristics. After the filtering, the speech samples were downsampled to 8 kHz, encoded and decoded with the AMR codec [27] and equalized to -26 dBov with SV56 [26, 28].

Because some of the Lombard samples in the training data did not have a clear Lombard effect, a better subset of the training data was selected using the SII [29]. An average SII score was computed for all normal and Lombard speech samples and the sample pairs where the Lombard speech had a higher score than the corresponding normal sample were chosen for the training. After this, the spectral tilts of the normal and Lombard samples were estimated with each of the estimation methods using 20 ms frames and transformed to line spectral frequency (LSF), reflection coefficient (RC) and log-area ratio (LAR) representations. To align the voiced frames of normal speech to the corresponding frames of Lombard speech, the dynamic time warping algorithm in [30] was utilized.

The statistical dependencies between the normal speech feature vectors $\boldsymbol{x}$ and the Lombard speech feature vectors $\boldsymbol{y}$ were modelled as a GMM

$$p(\boldsymbol{x},\boldsymbol{y}) = \sum_m w_m N\left(\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{\mu}_{x|m}\\ \boldsymbol{\mu}_{y|m}\end{bmatrix} \begin{bmatrix}\boldsymbol{\Sigma}_{xx|m} & \boldsymbol{\Sigma}_{xy|m}\\ \boldsymbol{\Sigma}_{yx|m} & \boldsymbol{\Sigma}_{xx|m}\end{bmatrix}\right), \quad (1)$$

where the component probabilities are denoted as w_m, the mean vectors as $\boldsymbol{\mu}_m$, and the covariance matrices as $\boldsymbol{\Sigma}_m$. Gaussian mixtures with $M = \{5, 10, 50, 100\}$ full-covariance components were considered. The model parameters were trained with the expectation-maximization algorithm implemented in [31].

The candidate models were evaluated in a Lombard tilt prediction task using the development data. Both the parameter representation (LP, LSF, RC or LAR) and the number of GMM components were varied across models. The minimum mean square error (MMSE) estimate for the features $\boldsymbol{y}(n)$ that correspond to $\boldsymbol{x}(n)$ was calculated based on the GMM distribution

$$\hat{\boldsymbol{y}}(n) = \sum_m P(m|\boldsymbol{x}(n)) \left[\boldsymbol{\mu}_{y|m} + \boldsymbol{A}_m(\boldsymbol{x}(n) - \boldsymbol{\mu}_{x|m})\right], \quad (2)$$

where the linear transformations $\boldsymbol{A}_m = \boldsymbol{\Sigma}_{yx|m}\boldsymbol{\Sigma}_{xx|m}^{-1}$ and the component probabilities $P(m|\boldsymbol{x}(n))$ were calculated based on the prior probabilities w_m and the feature likelihoods $N(\boldsymbol{x}(n)|\boldsymbol{\mu}_{x|m}, \boldsymbol{\Sigma}_{xx|m})$. The predicted features were compared to the ones calculated from the development data, and the best models were chosen based on the explained variance (R^2), calculated in the feature domain, and log-spectral distortion. If the 50 and 100 component models had similar R^2 values, the smaller, 50 component model was selected. The performance of the selected models was measured objectively in

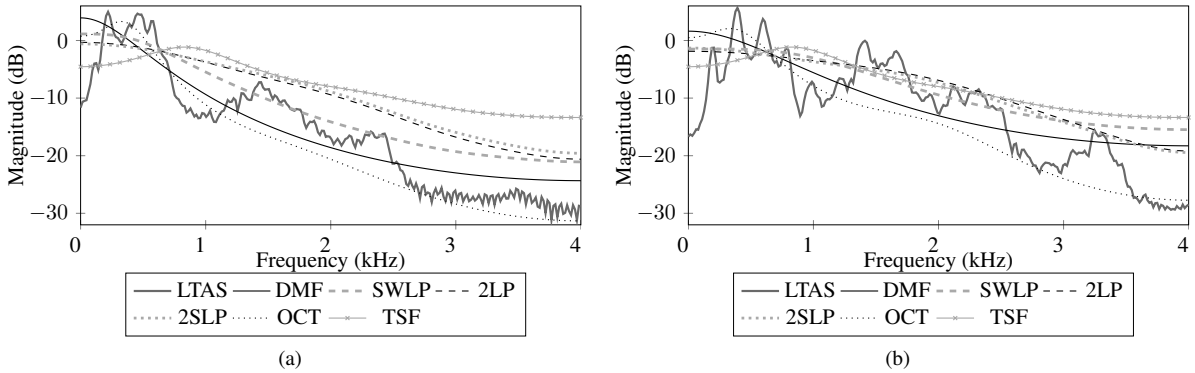

Figure 1: *The long-term average spectra of (a) normal and (b) Lombard speech and the estimated spectral envelopes of the methods under evaluation. The spectrum has been computed by averaging over all the voiced frames of a male speaker in the training database.*

Table 2: *Selected GMM model parameters (representation and number of GMM components, M), the explained variance (R^2) and the objective performance of the models for the spectral tilt estimation methods under comparison. The performance is measured in terms of frequency-weighted SNR (fwSNR, in dB) for which both the value achieved by the model (Achiev.) and the method's maximum value (Max.) are given.*

		DMF	SWLP	2LP	OCT	TSF
GMM	Repres.	RC	LSF	LSF	LSF	LSF
	M	5	10	10	50	50
	R^2	0.97	0.99	0.99	0.91	0.95
fwSNR	Achiev.	15.56	15.46	15.46	15.53	14.87
	Max.	15.87	16.30	16.10	16.51	15.35

terms of frequency-weighted SNR (fwSNR) [32] implemented as in [33]. The spectral tilt in the development data was replaced with the mapped tilt and the real Lombard speech was used as the target signal. However, differing from [33], the computation was done in the mel-spectrum domain utilizing 21 frequency bands. The model parameters selected for each of the spectral tilt estimation methods and their fwSNR values are shown in Table 2. The maximum obtainable fwSNR values in the table were computed by replacing the normal tilt directly with the estimated Lombard tilt. For comparison, the fwSNR value of unprocessed speech was 14.71 dB.

4. Post-processing algorithm

The incoming speech signal is processed with a 8-kHz sampling frequency in 20-ms frames which are first windowed with $w_n = \sin(\pi/(2N) \cdot (n + 0.5))$ [34], where N is the length of the window. The same window is also applied after the processing and 50 % overlap between consecutive frames is used. The energy and the gradient-index [35] are computed from the incoming speech frame, and used to classify the frame either as silence, unvoiced or voiced speech. Frames classified as silence or unvoiced are not processed. The flowchart of the processing for voiced frames is shown in Fig. 2.

First, the spectral tilt, denoted as $A_{\mathrm{p}}(z)$ in Fig. 2, is estimated utilizing one of the estimation methods, transformed to the desired representation (LSF or RC, depending on the method) and mapped with the corresponding GMM. After the mapping, the stability of the output filter, $A'_{\mathrm{p}}(z)$, is checked and if necessary, the roots outside of the unit circle are replaced with their mirror-image pairs inside the unit circle. The speech frame is then filtered with $H_{\mathrm{p}}(z) = A'_{\mathrm{p}}(z)/A_{\mathrm{p}}(z)$ (2LP) or $H_{\mathrm{p}}(z) = A_{\mathrm{p}}(z)/A'_{\mathrm{p}}(z)$ (other methods) which effectively removes the old spectral tilt and replaces it with the Lombard-like spectral tilt. Finally, the energy of the filtered frame is equalized to the level of the unprocessed frame with the adaptive gain control (AGC) found from the AMR codec [36].

5. Subjective evaluation

A subjective listening test was organized to evaluate the performance of the different tilt estimation methods in comparison to unprocessed speech (UN). The test had an intelligibility evaluation, in the form of a word-error rate (WER) test, followed by a pair comparison test on the quality of the samples. In the WER test, clean speech was corrupted with two types of additive noise (stationary car noise and unstationary factory noise [37]) both with two SNR levels which were selected based on informal listening to create noise conditions characterized as moderate, and severe. The SNR levels for car noise were -5 dB and -10 dB and for factory noise 0 dB and -5 dB. In the pair comparison test, only car noise with 20 dB SNR level was used.

The speech material consisted of phonetically balanced sentence material from two male and two female speakers which has been calibrated in terms of intelligibility in a previous study [38]. The speech material developed in [38] consists of meaningful sentences both in Finnish and English but only the Finnish material was used for the present study. The sentences contained 4-5 words and had an average duration of approximately 2 seconds. All speech samples were first preprocessed to correspond to narrowband telephone speech by utilizing the MSIN filter and AMR coding, as described in the second paragraph of Section 3. After this, the samples were modified with the post-processing algorithm, using each of the spectral tilt estimation methods under comparison, and finally, car or factory noise was added according to the noise condition.

Ten normal-hearing listeners, all native speakers of Finnish, participated in the listening tests. The tests took place in a sound-proofed listening booth using Sennheiser HD 650 headphones. The test was divided into three parts and a short practice session preceded each part. The A-weighted sound pressure level was set to 65 dB and kept constant throughout the tests.

In the WER test, each sample was played only once after which the subjects typed the sentence on the computer. The percentage of correct words was computed by scoring the stems and suffices of inflected words separately after obvious spelling

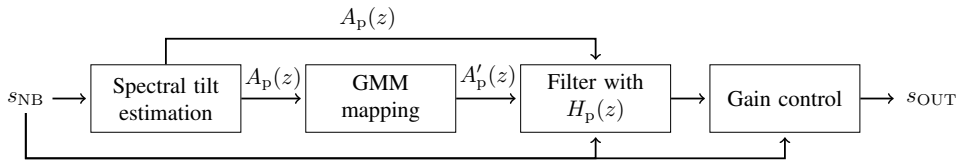

Figure 2: *The flowchart of the post-processing algorithm for voiced frames. The incoming speech frame is denoted by s_{NB} and the processed speech frame by s_{OUT}.*

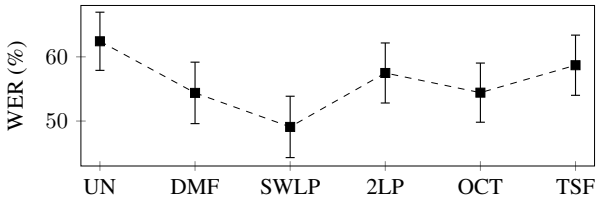

Figure 3: *The average word-error rates (WER) and their 95 % confidence intervals for unprocessed speech (UN) and the five spectral tilt estimation methods under comparison (DMF, SWLP, 2LP, OCT, TSF). The results have been averaged over the noise conditions and speakers.*

errors had been corrected. In the pair comparison test, the listeners were able to freely listen to two samples, A and B, and were asked *"Which sample is of better quality?"*. They were asked to choose one of the options: *A*, *B* or *No difference* and instructed to select *No difference* if they had no preference even if they heard a difference between the samples.

5.1. Results

The results of the WER test were analysed with a five-way analysis of variance (ANOVA) procedure using 5% significance level. The test subject was modelled as a random factor while the noise type (car, factory), the SNR level (moderate, severe), the method (UN, DMF, SWLP, 2LP, OCT, TSF), and the speaker gender (male, female) were modelled as fixed factors. Although a mild violation of the underlying ANOVA assumption about the normality of the residuals was discovered in visual inspection of the data, the ANOVA procedure was considered applicable due to the robustness of such linear mixed effects models [39]. The ANOVA revealed that the noise type [F(1,14) = 6.77, $p < 0.05$], the method [F(5,45) =5.11, $p < 0.001$], the SNR level [F(1,9) = 531.67, $p < 0.001$], the speaker gender [F(1,9) = 199.21, $p < 0.001$] as well as the interaction between the noise type, the SNR level and the speaker gender [F(1,9) = 12.48, $p < 0.05$] had a significant effect on the WER scores.

The next step in the analysis comprised computation of the marginal means and their 95% confidence intervals in order to gain detailed knowledge about the nature of the effects. The Dunnett's T3 post-hoc test with 5% significance level was applied to confirm the statistical significance of the findings. The values shown in Fig. 3 illustrate that only SWLP was able to reduce the WER, on average, as compared to UN. DMF and OCT also seem to yield reduction of WER, but their difference to the scores achieved with UN failed to reach statistical significance.

The responses given by the test subjects in the paired comparison test were encoded into two separate preference matrices, one for each speaker gender. After this, the Bradley-Terry method was used to fit generalized linear models to the data obtained, and the two-way ANOVA procedure was used to test whether the preference score depended on the method or on the gender of the speaker. The comparison type [χ^2 = 97.006, d.f. = 5, $p < 0.001$] had a significant effect on whether the participant would select one of the compared samples.

Table 3: *Pairwise comparison between methods in terms of overall quality. Only such comparisons where significant preference was found are shown. The preferred method is highlighted with the letters in boldface.*

Comparison	W	p
UN-OCT	-4.80	0.00
DMF-OCT	-5.38	0.00
SWLP-OCT	-5.72	0.00
2LP-OCT	-5.09	0.00
TSF-OCT	-5.38	0.00

The paired comparisons were further analysed in a pairwise manner to obtain detailed knowledge on whether a particular method was preferred significantly over another method. These analyses were performed using the Barnard's exact test with 5% significance level. Table 3 summarizes the results for the comparisons in which one of the method was significantly preferred over the other. Inspection of the results reveals that all other methods were preferred over OCT, while no significant differences were found between UN and DMF, SWLP, 2LP or TSF.

6. Conclusion

A GMM-based post-processing method, targeted for intelligibility enhancement in transmission of telephone speech, was proposed. Six spectral tilt estimation methods (DMF, SWLP, 2LP, 2SLP, OCT, TSF) were evaluated for the purpose of statistical mapping from normal to Lombard speech. Based on an initial objective evaluation of the tilt estimation methods, one of them, 2SLP, was removed from further evaluations. GMMs were trained for the five remaining methods and their performance in post-processing was evaluated objectively as well as in subjective listening tests in terms of intelligibility and quality compared to unprocessed speech. The results indicate that SWLP was the only method able to increase the intelligibility of speech. However, considering the fairly mild Lombard effect in the training data and the averaging tendency of the GMM, the result is encouraging. OCT and DMF also showed slight indication of intelligibility improvement but OCT was also rated the lowest in quality. This is clearly a problem of the estimation method itself and does not relate to the GMM mapping. However, future work will also include the evaluation of alternative statistical mappings.

7. Acknowledgements

The work was supported by the GETA graduate school, Nokia, the Academy of Finland (proj. no. 256961, 13251770, 136209, 251170), and the Mide/Ui-art project of Aalto University.

8. References

[1] B. Sauert and P. Vary, "Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement," in *ITG-Fachtagung Sprachkommunikation*, 2010.

[2] C. Taal, J. Jensen, and A. Leijon, "On optimal linear filtering of speech for near-end listening enhancement." *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 225–228, 2013.

[3] H. Schepker, J. Rennies, and S. Doclo, "Improving speech intelligibility in noise by SII-dependent preprocessing using frequency-dependent amplification and dynamic range compression," in *Proc. Interspeech*, 2013, pp. 3577–3581.

[4] Y. Tang and M. Cooke, "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *Proc. Interspeech*, 2012.

[5] R. Niederjohn and J. Grotelueschen, "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, pp. 277–282, 1976.

[6] J. Hall and J. Flanagan, "Intelligibility and listener preference of telephone speech in the presence of babble noise," *J. Acoust. Soc. Amer.*, vol. 127, no. 1, pp. 280–285, 2010.

[7] E. Jokinen, S. Yrttiaho, H. Pulakka, M. Vainio, and P. Alku, "Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech," *J. Acoust. Soc. Amer.*, vol. 132, no. 6, pp. 3990–4001, 2012.

[8] W. V. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 917–928, 1988.

[9] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Commun.*, vol. 51, no. 12, pp. 1253–1262, 2009.

[10] M. Cooke and Y. Lu, "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 2059–2069, 2010.

[11] M. Skowronski and J. Harris, "Applied principles of clear and Lombard speech for automated intelligibility enhancement in noisy environments," *Speech Commun.*, vol. 48, no. 5, pp. 549–558, 2006.

[12] T.-C. Zorilă, V. Kandia, and Y. Stylianou, "Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression," in *Proc. Interspeech*, 2012.

[13] E. Jokinen, P. Alku, and M. Vainio, "Lombard-motivated post-filtering method for the intelligibility enhancement of telephone speech," in *Proc. Interspeech*, 2012.

[14] E. Jokinen, M. Takanen, M. Vainio, and P. Alku, "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Comput., speech, lang.*, vol. 28, no. 2, pp. 619–628, 2014.

[15] E. Godoy, M. Koutsogiannaki, and Y. Stylianou, "Approaching speech intelligibility enhancement with inspiration from Lombard and clear speaking styles," *Comput., speech, lang.*, vol. 28, no. 2, pp. 629–647, 2014.

[16] C. Valentini-Botinhao, J. Yamagishi, S. King, and Y. Stylianou, "Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise," in *Proc. Interspeech*, 2013, pp. 3567–3571.

[17] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise," *Comput., Speech, Lang.*, vol. 28, no. 2, pp. 648–664, 2014.

[18] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech, Audio Process.*, vol. 6, no. 2, pp. 131–142, 1998.

[19] H. Mizuno and M. Abe, "Voice conversion algorithm based on piecewise linear conversion rules of formant frequency and spectrum tilt," *Speech Commun.*, vol. 16, no. 2, pp. 153–164, 1995.

[20] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Commun.*, vol. 51, no. 5, pp. 401–411, 2009.

[21] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Commun.*, vol. 12, no. 1, pp. 69–81, 1993.

[22] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, no. 3, pp. 283–296, 1975.

[23] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 873–881, 2007.

[24] J. Gray, A. and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 5, pp. 380–391, 1976.

[25] M. Vainio, D. Aalto, A. Suni, A. Arnhold, T. Raitio, H. Seijo, J. Järvikivi, and P. Alku, "Effect of noise type and level on focus related fundamental frequency changes," in *Proc. Interspeech*, 2012.

[26] *Recommendation G.191: Software tools for speech and audio coding standardization*, Int. Telecommun. Union, Geneva, Switzerland, September 2005.

[27] *Specification TS 26.104: ANSI-C code for the floating-point Adaptive Multi-Rate (AMR) speech codec*, 3rd Generation Partnership Project, Valbonne, France, 2009, version 9.0.0.

[28] *Recommendation P.56: Objective measurement of active speech level*, Int. Telecommun. Union, Geneva, Switzerland, March 1993.

[29] *American national standard ANSI S3.5-1997: Methods for calculation of the speech intelligibility index*, American national standards institute, Inc., 1997.

[30] D. Ellis. (2003) Dynamic time warp (DTW) in Matlab. Visited 16.03.2014. [Online]. Available: http://www.ee.columbia.edu/ dpwe/resources/matlab/dtw/

[31] P. Paalanen, J. Kämäräinen, and H. Kälviäinen. (2005) GMMBayes toolbox for Matlab - Gaussian mixture model learning and Bayesian classification. Visited 22.03.2014. [Online]. Available: http://www.it.lut.fi/project/gmmbayes/

[32] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 1978, pp. 586–590.

[33] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[34] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

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

[36] *Specification TS 26.090: Adaptive multi-rate (AMR) speech codec; Transcoding functions*, 3rd Generation Partnership Project, Valbonne, France, 2008, version 8.0.0.

[37] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[38] M. Vainio, A. Suni, H. Järveläinen, J. Järvikivi, and V.-V. Mattila, "Developing a speech intelligibility test based on measuring speech reception thresholds in noise for English and Finnish," *J. Acoust. Soc. Amer.*, vol. 118, no. 3, pp. 1742–1750, 2005.

[39] H. Jaqmin-Gadda, "Robustness of the linear mixed model to misspecified error distribution," *Comput. Stat. Data Analysis*, vol. 51, no. 10, pp. 5142–5154, 2007.