INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The use of read versus conversational Lombard speech in spectral tilt modeling for intelligibility enhancement in near-end noise conditions

Emma Jokinen, Ulpu Remes and Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
emma.jokinen@aalto.fi

Abstract

Intelligibility of speech in adverse near-end noise conditions can be enhanced with post-processing. Recently, a post-processing method based on statistical mapping of the spectral tilt of normal speech to that of Lombard speech was proposed. However, previous intelligibility improvement studies utilizing Lombard speech have mainly gathered data from read sentences which might result in a less pronounced Lombard effect. Having a mild Lombard effect in the training data weakens the statistical normal-to-Lombard mapping of the spectral tilt which in turn deteriorates performance of intelligibility enhancement. Therefore, a database containing both conversational and read Lombard speech was recorded in several background noise conditions in this study. Statistical models for normal-to-Lombard mapping of the spectral tilt were then trained using the obtained conversational and read speech data and evaluated using an objective intelligibility metric. The results suggest that the conversational data contains a more pronounced Lombard effect and could be used to obtain better statistical models for intelligibility enhancement.

Index Terms: Lombard speech, conversational recording, intelligibility enhancement, Gaussian mixture model

1. Introduction

In mobile communications, the quality and intelligibility of degraded speech can be enhanced with post-processing. This degradation can be the result of quantization or acoustical background noise on the sending or receiving side of the connection, referred to as the far-end and the near-end noise scenario, respectively. In this study, the decoded speech is assumed to contain only quantization noise and the noise disrupting the communication is in the listener's environment. In this scenario, the post-processing is used to enhance the acoustic cues in the clean speech to improve its intelligibility over the background noise.

Several intelligibility enhancement methods have been proposed previously for the near-end noise scenario. They are based, for example, on optimizing objective measures [1, 2, 3, 4] or re-allocating speech energy with simple high-pass filtering [5, 6]. While most of these techniques are based on models of human speech perception, methods based on modeling the human speech *production* mechanism are not as common. It is well known that speakers tend to change, for example, the spectral characteristics of their speech when talking in noisy conditions. An example of this is the Lombard effect which refers to the modification of speaking style due to environmental noise [7]. Natural Lombard speech has been shown to be more intelligible than normal speech and this has been attributed to several factors, such as flattening of the spectral tilt, slower speaking rate, and increased vocal intensity [8, 9]. The Lombard effect has been taken advantage of in some previous post-processing studies [10, 11, 12], and recently, Gaussian mixture models (GMMs) were used in a normal-to-Lombard mapping successfully to improve the intelligibility of telephone speech [13, 14].

Previous intelligibility enhancement studies utilizing the Lombard effect are generally based on speech recordings in which the speaker reads pre-selected sentences in noise [12, 13]. This scenario is different from a realistic speech communication situation between two talkers in noisy conditions in which both interlocutors spontaneously adjust their speaking style to tackle the disturbance caused by noise to deliver the spoken message to the other partner. In other words, the importance of person-to-person communication, the concept referred to as communicative interaction [15] or intent [16], has unfortunately been ignored in data collection in previous intelligibility enhancement studies. This is regrettable because some studies [15, 17, 18] have shown that the interaction seems to evoke a stronger Lombard effect compared to reading as observed, for example, in increased vocal intensity [15]. From the point of view of machine learning, having a more pronounced Lombard effect in the training data might lead to statistical models better suited for post-processing in severe background noise conditions. Previously, different kind of data selection techniques have been used in an effort to obtain representative Lombard data [19].

In this study, a database containing both conversational and read Lombard speech is recorded in several background noise conditions. Both the obtained conversational and read speech data are then used to train statistical models for normal-to-Lombard mapping. Finally, the models are used in a post-processing context and evaluated using an objective intelligibility metric, the speech intelligibility index [20].

2. Lombard recordings

A database of conversational and read Lombard speech in several different noise conditions was recorded from several Finnish speakers. Each recording session consisted of two parts: in the first part, realistic communication tasks were used to create an interaction between two subjects. Each pair of interlocutors completed several different tasks, one for each background noise condition. In the second part, each individual read through a short text once in each background noise condition.

To obtain conversational speech data, a task evoking interaction between subjects is needed. Several different kinds of tasks have been proposed in previous studies, such as, interactive map drawing [21, 15] and locating the differences in pictures [22, 16]. In the current study, realistic telephone conversations were generated with communication tasks designed for conversational quality evaluation of telephone connections [23]. In the tasks, one of the subjects is the caller and the other one is

http://dx.doi.org/10.21437/Interspeech.2016-143

Figure 1: The test setup of the conversational tasks in the anechoic chamber. The controller of the experiment was seated where the picture was taken from.

an agent in a service, such as travel agency or library. Each subject has a task sheet, which, for the caller, specifies the service he/she is asking for, and for the agent tells what kind of service options are available. Each task sheet is missing information that the other task sheet contains, hence generating spontaneous interaction to exchange the necessary information. A part of the tasks suggested in the standard have been previously translated to Finnish for the evaluation of bandwidth extension [24] and these translated versions were adapted for the current study. Altogether six conversation tasks were used: one for a practice session and five others for different background noise conditions. The tasks included buying train tickets, reserving plane tickets, booking seats for the theater, buying tickets to a musical and reserving books from a library. The time to complete one task varied around 3 minutes. In the second part of the data collection, participants read a text (weather forecast) of 90 words which took approximately one minute.

Altogether five background noise conditions were used for each pair of interlocutors: silence, stationary car noise [25] as well as highly unstationary pub noise [25], both with A-weighted sound pressure levels (SPL) of approximately 65 dB and 80 dB. Previous studies describing the collection of Lombard data were used as reference for selecting the appropriate noise levels [8, 26, 27]. The presentation order of the background noises was kept fixed for all pairs of subjects. The noise was played to the subjects using Sennheiser HD 595 headphones that were set to the same sound pressure level. The open headphones also allowed the subjects to hear their own voice to some extent. Both of the subjects had a headset microphone (DPA 4065-BL and DPA 4066-B) as well as a stand microphone (G.R.A.S. 46AF 1/2" free-field microphone) that were used for recording. The recording for each participant was done on two channels with different amplifications in order to avoid clipping of the recorded signals. The signal from the standing microphone was routed to the other subjects headphones so that the resulting SPL in the ear was approximately 55 dBA in silence. All the mixing was done using a MOTU Ultralite-mk3 Hybrid audio interface that was connected to a computer via USB. The signals were recorded at 48-kHz sampling frequency using the REAPER software. In addition to the speech data, a calibration signal, a 1-kHz sine tone, was recorded using all the microphones so that the speech SPL could be determined afterwards from the digital signals.

	Conversational					Read				
	1	2	3	4	5	1	2	3	4	5
M	77	84	86	85	87	69	72	73	72	73
F	80	86	86	86	87	71	73	74	73	75

Table 1: The average A-weighted sound pressure levels (in dB) computed from the two male (M) and two female (F) speakers analyzed in the current study. The values are calculated for both conversational and read speech in all the background noise conditions (1 = silence, 2 = moderate car, 3 = severe car, 4 = moderate pub, 5 = severe pub).

Ten pairs of subjects took part in the recordings, altogether 10 male and 10 female speakers. All of the participants were either native or bilingual speakers of Finnish. The average age of the speakers was 23 years and they were either university students or staff. The subjects were paid for their participation. The recordings were conducted in an anechoic chamber where the subjects were seated on the opposite sides of the chamber as shown in Fig. 1. In between them was a curtain blocking the view so the subjects could not communicate using gestures. Each of the participants had a wooden clip board that they could use to write the required information to their task sheets. In the beginning of the recording session, the participants were instructed on the recording procedure and on how to complete the tasks. First, a practice task was completed with silence in the background followed by the other five tasks in different noise conditions. Between each task, the subjects were able to relax, drink water and study the task sheet for the next task. After completing the conversational tasks, each participant conducted the text reading part while the other one waited outside of the anechoic chamber.

For this preliminary study, four speakers (two male, two female) were selected from the recorded dataset. The average A-weighted SPL values computed from the selected speakers in all the different background noise conditions are shown in Table 1. The average values in the conversational conditions are generally higher than in the read conditions and the increase in SPL from speech produced in silence to speech produced in noise is slightly larger in the conversational data.

3. Normal-to-Lombard conversion

The main motivation of recording conversational Lombard speech was to obtain more representative speech data for training a normal-to-Lombard mapping compared to read speech. In order to verify this, Gaussian mixture models for converting the spectral tilt of normal speech to that of Lombard speech were trained using both the conversational and read data. The entire process was first introduced in [13]. Although the recordings contain Lombard speech data in four noisy conditions, only one of them, the severe pub noise condition, was selected for GMM training in the present study. Speech produced in this condition will simply be referred to as Lombard speech (L), whereas speech produced with silence in the background is referred to as normal speech (N). As mentioned previously, a subset of two male and two female speakers was used for training the models in the current investigation.

First, both the conversational and the read recordings were divided into smaller parts and the unnecessary silences were removed. The conversational recordings contain also some non-speech sounds that are often used to indicate, for instance, hesitation or agreement. Most of these were also removed at the

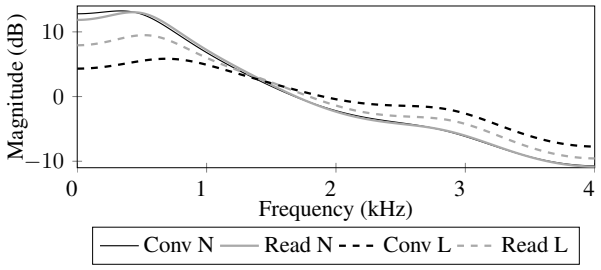

Figure 2: Average spectral tilt for both conversational and read normal (N) and Lombard (L) speech for a male speaker. The spectral tilt has been computed from all the voiced frames using stabilized-weighted linear prediction.

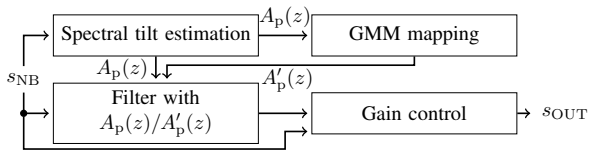

Figure 3: The flowchart of the post-processing algorithm for voiced frames. The incoming speech frame is denoted by s_{NB} and the processed speech frame by s_{OUT}.

pre-processing stage. After this, all the speech data was downsampled from 48 kHz to 16 kHz. Because the main interest of the current study was the effects of spectral tilt on intelligibility, all the level differences between different types of speech were removed. This was achieved by equalizing the signals to -26 dBov using SV56 [28, 29]. After this, all the speech samples were downsampled to 8 kHz. The read normal and Lombard datasets contain parallel recordings, that is to say, the same sentences were produced by each speaker using normal and Lombard speaking styles. To find the corresponding normal and Lombard frames from the read data, dynamic time warping (DTW) [30] was used with 20-ms frames. After this, the voiced frames were selected using both the energy of the frame as well as the gradient-index measure [31].

Whereas the read normal and Lombard data are parallel, the conversational normal and Lombard speech have different linguistic content. Finding matching frames for training a statistical mapping is, therefore, not as straightforward as with the read speech data. An additional challenge in the current study is the large difference between normal and Lombard speech. This difference, while highly desirable for intelligibility enhancement, complicates finding correct matches between normal and Lombard data. However, in the current study, the dataset contains also parallel, read speech data from the same speakers produced in the same noise conditions which was used to match the normal and Lombard frames in conversational speech.

First, the voiced frames were selected from the conversational data with the same method applied previously to the read frames and the mel-frequency cepstral coefficients (MFCC) were computed for both the read and the conversational data using a frame length of 20 ms. The length of the MFCC feature vector was 12 (the 0th coefficient was excluded). These features were then used to find nearest neighbors in terms of the Euclidean distance for the conversational frames from the corresponding read frames of the same speaker. In other words, the conversational normal frames were compared to the read

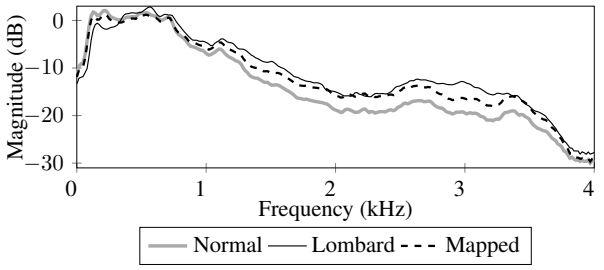

Figure 4: Long-term average spectra (LTAS) for a male speaker computed for the read speech data. The LTAS of both the read normal and the read Lombard data along with the speech processed with the GMM mapping trained on read speech are shown. The processed speech is referred to as Mapped in the figure.

normal frames and similarly for the Lombard frames. If a conversational normal frame and a conversational Lombard frame were associated with normal and Lombard frames from the read data that had been aligned by the DTW, the conversational normal and Lombard frame were selected for the GMM training. If multiple conversational frames were matched to one read frame, the conversational frame with the smallest Euclidean distance in the MFCC feature space was selected. Using this procedure, 6091 conversational frames were selected for GMM training. The number of frames in the read training data was 9670.

For the statistical normal-to-Lombard mapping, the spectral envelope of speech needs to expressed parametrically. In this study, stabilized weighted linear prediction (SWLP) was used to parametrize the tilt of the voiced frames. SWLP [32] is an all-pole modeling technique similar to weighted linear prediction (WLP) [33] in which the square of the residual is temporally weighted based on the short-time energy (STE) of the speech signal. SWLP has been used previously in [13] for the mapping of spectral tilt with GMMs. Optimization of the SWLP parameters in [14] yielded values $M = 2$ and $p = 6$ which were used throughout this study. The final feature vector contained the SWLP features as line spectral frequencies (LSFs). Examples of the average spectral tilt obtained using the SWLP parametrization for a male speaker are shown in Fig. 2. While the average spectral tilt computed from the conversational speech is quite close to the one computed from read speech, the difference in the tilt between conversational Lombard and read Lombard speech is notable.

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

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

where the component probabilities are denoted as w_i, the mean vectors as $\boldsymbol{\mu}_i$, and the covariance matrices as $\boldsymbol{\Sigma}_i$. The model parameters are trained with the expectation-maximization algorithm implemented in [34]. The minimum mean square error (MMSE) estimate for features $\boldsymbol{y}^*$ that correspond to test input $\boldsymbol{x}^*$ is calculated based on the GMM distribution as

$$\boldsymbol{y}^* = \sum_i P(i|\boldsymbol{x}^*) \left[\boldsymbol{\mu}_{y|i} + \boldsymbol{A}_i(\boldsymbol{x}^* - \boldsymbol{\mu}_{x|i}) \right], \quad (2)$$

where the linear transformations $\boldsymbol{A}_i = \boldsymbol{\Sigma}_{yx|i}\boldsymbol{\Sigma}_{xx|i}^{-1}$ and the component probabilities $P(i|\boldsymbol{x}^*)$ are calculated based

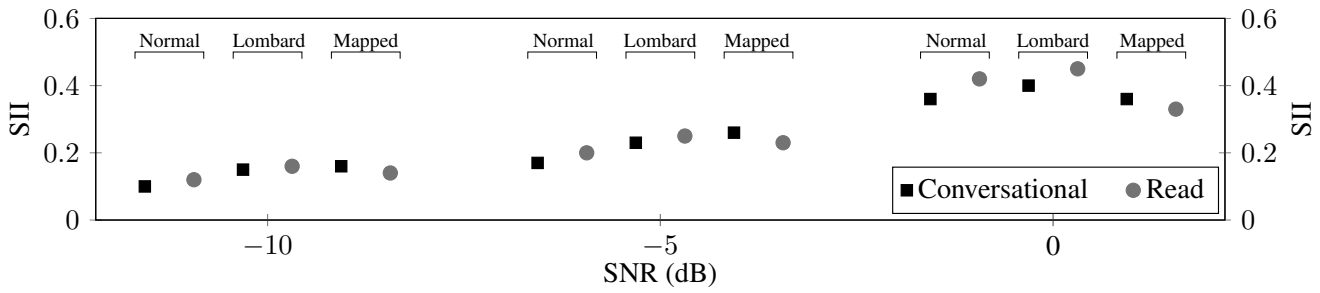

Figure 5: The objective intelligibility scores measured using the speech intelligibility index (SII) for both the original conversational and read normal and Lombard datasets as well as for speech processed using both the GMM mapping trained with conversational data and the one trained with read data. The processed speech is referred to as Mapped in the figure. The results are averaged for all the speakers in the evaluation.

on the prior probabilities w_i and the feature likelihoods $N(\boldsymbol{x}^*|\boldsymbol{\mu}_{x|i}, \boldsymbol{\Sigma}_{xx|i})$. GMM models with $I = 5$ full-covariance components were used in the current study.

The mapping of spectral tilt from normal to Lombard speech is utilized as a part of a post-processing algorithm which has been previously used in [13]. The flowchart of the processing for voiced frames is shown in Fig. 3. The incoming speech signal is processed with a 8-kHz sampling frequency in 20-ms frames which are first windowed with $w_l = \sin(\pi/(2L) \cdot (l + 0.5))$ [35], where L is the length of the window. The same window is also applied after the processing with 50 % overlap between consecutive frames. The energy and the gradient-index [31] are computed from the incoming speech frame, and used to classify the frame. Frames classified as silence or unvoiced are not processed.

First, the spectral tilt, parametrized as $1/A_{\mathrm{p}}(z)$ in Fig. 3, is estimated with SWLP, transformed to the LSF representation and mapped with the trained model. After the mapping, the stability of the output filter is checked and if necessary, the roots outside of the unit circle are replaced with their mirror-image pairs inside the unit circle. The frame is then filtered with $A_{\mathrm{p}}(z)/A'_{\mathrm{p}}(z)$ removing the original spectral tilt and replacing 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) in the AMR codec [36].

4. Objective evaluation

The performance of the normal-to-Lombard mapping based post-processing methods trained on conversational and read speech were evaluated using the speech intelligibility index (SII) [20]. The speech data used for the testing was the read normal speech of the same four speakers. A visualization of the long-term average spectra (LTAS) computed from speech of a male speaker is shown in Fig. 4.

In addition to the processed samples, the objective intelligibility values were also computed for the original normal and Lombard data for both the conversational and read case. As mentioned previously, all the samples were normalized to the same level using SV56 [29, 28]. Stationary car noise with three different SNR levels, -10 dB, -5 dB, and 0 dB, was used as background noise in the evaluation. The SII metric was computed for each speech sample by first removing silent periods from the samples, computing the SII in segments of 9.4 ms and then averaging the obtained values over the whole sample. All the values were then averaged over the four speakers. The resulting SII values are shown in Fig. 5.

Interestingly, the equalized conversational data shows consistently slightly lower SII scores than the corresponding read speech data. The conversational data still contains segments where the subjects are mostly talking to themselves while they are taking notes which might in part be behind the lower intelligibility scores. However, the difference between the normal and Lombard cases is larger in the conversational data than in the read data. The processed speech receives similar scores as the read Lombard speech in the lowest SNRs and the mapping trained with conversational speech yields even slightly higher intelligibility scores than the real Lombard speech. Notably, the baseline for the processed speech is the read normal speech, not the conversational normal speech.

5. Discussion

A database containing both conversational and read Lombard speech was recorded in several background noise conditions. A separate statistical normal-to-Lombard mapping for the spectral tilt was then trained using both the obtained conversational and read speech data. The normal-to-Lombard conversions were then evaluated and compared to the original speech data using the speech intelligibility index.

The results suggest that the conversational Lombard data presents a more pronounced effect and, thus, might be better suited for training statistical models for intelligibility enhancement. However, while the normal-to-Lombard conversion has a clear impact on both the LTAS and the average spectral tilt, the difference between the two types of processed speech is much smaller than the difference between the original data shown in Fig. 2. A similar tendency can be observed in the objective intelligibility scores. One major factor in this disability to transfer the effect in the conversational data to the statistical mapping might be the selection of training data. The problem of using non-parallel data for voice conversion has been discussed, for instance, in [37, 38]. The data selection technique used in the current study might favor conversational frames that resemble read frames which would result in similar statistical models for the two datasets.

6. Acknowledgement

This work was supported by the Academy of Finland (projects 256961, 284671 and 269279). The authors would like to thank Ilkka Huhtakallio (MSc) for help in setting up the recordings.

7. 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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] E. Jokinen, U. Remes, M. Takanen, K. Palomäki, M. Kurimo, and P. Alku, "Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech," in *Proc. Int. Workshop Acoust. Signal Enh. (IWAENC)*, 2014, pp. 164–168.

[14] ——, "Spectral tilt modelling with GMMs for intelligibility enhancement of narrowband telephone speech," in *Proc. Interspeech*, 2014, pp. 2036–2040.

[15] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the Lombard effect," *J. Speech, Lang., Hear. Res.*, vol. 53, no. 3, pp. 588–608, 2010.

[16] R. Baker and V. Hazan, "LUCID: a corpus of spontaneous and read clear speech in British English," in *Proc. DiSS-LPSS Joint Workshop 2010*, 2010, pp. 1–4.

[17] D. K. Amazi and S. R. Garber, "The Lombard sign as a function of age and task," *J. Speech, Lang., Hear. Res.*, vol. 25, no. 4, pp. 581–585, 1982.

[18] J.-C. Junqua, S. Fincke, and K. Field, "Influence of the speaking style and the noise spectral tilt on the Lombard reflex and automatic speech recognition," in *Proc. Int. Conf. Spoken Lang. Proc. (ICSLP)*, 1998, pp. 467–470.

[19] E. Jokinen, U. Remes, and P. Alku, "Comparison of Gaussian process regression and Gaussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech," in *Proc. Interspeech*, 2015, pp. 85–89.

[20] *American National Standard ANSI S3.5-1997: Methods for calculation of the speech intelligibility index*, American National Standards Institute, Inc., 1997.

[21] A. Anderson, M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," *Lang. Speech*, vol. 34, pp. 351–366, 1991.

[22] K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow, "The Wildcat corpus of native-and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles," *Lang. Speech*, vol. 53, no. 4, pp. 510–540, 2010.

[23] *Recommendation P.805: Subjective evaluation of conversational quality*, International Telecommunication Union, Geneva, Switzerland, April 2007.

[24] H. Pulakka, L. Laaksonen, S. Yrttiaho, V. Myllylä, and P. Alku, "Conversational quality evaluation of artificial bandwidth extension of telephone speech," *J. Acoust. Soc. Amer.*, vol. 132, no. 2, pp. 848–861, 2012.

[25] *Speech and multimedia Transmission Quality (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database*, ETSI, Sophia Antipolis Cedex, France, 2011, version 1.2.4.

[26] J. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 366–378, 2009.

[27] 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.

[28] *Recommendation G.191: Software tools for speech and audio coding standardization*, International Telecommunication Union, Geneva, Switzerland, September 2005.

[29] *Recommendation P.56: Objective measurement of active speech level*, International Telecommunication Union, Geneva, Switzerland, March 1993.

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

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

[33] 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.

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

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

[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] C.-H. Lee and C.-H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training." in *Proc. Interspeech*, 2006.

[38] D. Erro, A. Moreno, and A. Bonafonte, "INCA algorithm for training voice conversion systems from nonparallel corpora," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 944–953, 2010.