A Study on Combined Effects of Reverberation and Increased Vocal Effort on ASR

Hynek Bořil, Seyed Omid Sadjadi, and John H.L. Hansen

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, U.S.A.
{hynek, sadjadi, john.hansen}@utdallas.edu

Abstract

This study analyzes the individual and combined effect of room reverberation and increased vocal effort on automatic speech recognition. Robustness of several state-of-the-art front-end feature extraction strategies and normalizations to these sources of speech signal variability is evaluated in the context of large and small vocabulary recognition tasks on American English and Czech speech corpora. For the large vocabulary task, speech material from the UT-Scope database comprising American English utterances is used. The Czech speech samples are drawn from the CLSD‘05 data corpus and used for the small vocabulary tasks. Both databases contain neutral as well as increased vocal effort recordings. Simulated reverberant test conditions are generated using measured room impulse responses from the AIR database and utilized in the evaluations. It is shown that the robustness of a common MFCC front-end to reverberation and increased vocal effort can be considerably improved when paired with cepstral gain normalization and modified RASTA filtering. A combination of recently proposed mean Hilbert envelope coefficients and modified RASTA is found to provide balanced performance across all reverberation and vocal effort conditions.

1. Introduction

Room reverberation can cause various destructive impacts on spectro-temporal characteristics of speech signals, most notably including temporal smearing, filling dips and gaps in the temporal envelope, increasing the prominence of low-frequency energy, and flattening the formant transitions. These impacts have been categorized as self- and overlap-masking effects [1]. The self-masking effect is caused by early sound reflections in the room that arrive at the receiver (ear or microphone) within 50-80 ms after the direct sound. The overlap-masking effect on the other hand is resulted from late echos (or reflections) which tend to smear the direct sound over time and mask succeeding sounds. It has been shown that the overlap-masking effect of reverberation is the primary cause of degraded speech recognition performance in both human listeners [1] and automatic speech recognizers [2, 3].

In addition to signal distortion, room reverberation may result in increased vocal effort of the speakers [4]. This is due to the fact that room reverberation decreases speech quality and intelligibility, which in turn induces changes in the auditory feedback process. Consequently, speakers increase their vocal effort to compensate for the drop in intelligibility. This increase in vocal effort, which is a function of both reverberation time (aka T_{60}) and talker-to-listener distance [4], has been shown to be a major source of speech signal variability that can ultimately deteriorate performance of ASR.

Hence, in a reverberant environment, an ASR system has to struggle with not only the signal distortions, but also the signal variability due to the increased vocal effort which is induced by reverberation. There have been several research attempts that considered individual impacts of room reverberation [2, 3, 5–7] and increased vocal effort [8, 9] on ASR, and reported compensation strategies to alleviate these impacts. However, to the best of our knowledge, this study is one of the first to consider the individual as well as the combined effects of reverberation and increased vocal effort on ASR. In addition, robustness of various conventional and recently proposed feature extraction/compensation techniques are evaluated in the context of both small and large vocabulary ASR tasks under reverberation, increased vocal effort, and their combination. In particular, motivated by their encouraging performance in speaker identification (SID) under reverberation, the recently proposed mean Hilbert envelope coefficient (MHEC) features [10] are benchmarked against traditional MFCC preceded by long-term log spectral subtraction (LTLSS) [3] and Gammatone subband based non-negative matrix factorization (NMF) [7], as well as MFCC implemented in ETSI advanced front-end (AFE) [11], in our ASR experiments. The feature extraction schemes are paired with a number of popular cepstral normalizations and also recently proposed RASTALP temporal filtering.

2. Mean Hilbert Envelope Coefficients: MHEC

MHEC features have been shown to be an effective alternative to MFCCs for robust SID and ASR tasks under reverberant mismatched conditions [10, 12]. Here, we briefly describe the procedure for MHEC extraction.

First, the pre-emphasized reverberant speech signal is analyzed through a 26-channel Gammatone filterbank. Next, since we are mostly interested in slowly varying amplitude modulations rather than the fine structure, in each channel the Hilbert envelope is calculated and smoothed using a low-pass filter with a cut-off frequency of 20 Hz. In the next stage, the low-pass filtered envelope is blocked into frames of 25 ms duration with a skip rate of 10 ms. To estimate the temporal envelope amplitude in each frame, the sample mean is computed. Note that the sample mean is a measure of the spectral energy at the center frequency of each channel, and therefore overall provides a short-term spectral representation of the speech signal. Next, in each channel, the envelope trajectories are normalized using the long-term average computed over the entire utterance. This stage, which is called subband normalization (SN), functions as an automatic gain control (AGC) and is used to suppress any spectral coloration effect of the reverberation (or the self-masking effect) in different frequency channels. Up to this stage, only the self-masking effect which is due to early reflections has been suppressed. The overlap-masking effect, which is the long-term effect of reverberation and due to late reflections, can be modeled as an uncorrelated additive noise [6], and hence can be compensated via spectral subtraction [13]. The output of this stage represents an estimate of the clean anechoic speech spectrum. In the last stage, natural logarithm is applied to compress the dynamic range of spectral coefficients and followed by the DCT to obtain cepstral features. Here, only the first 13 coefficients (including c_0) are retained after DCT. The final output is a matrix of 13-dimensional cepstral features, entitled the mean Hilbert envelope coefficients (MHEC).

3. Feature Normalizations

Feature normalizations are typically used to transform incoming signal towards characteristics learned by the acoustic models. Depending on the type of normalization, detrimental effects of environmental acoustics, ambient noise, channel variations, as well as variations introduced by speakers can be addressed. In our study, several normalizations that were reported to increase robustness to channel variations and increased vocal effort are evaluated. It is noted that room reverberation can be viewed as a form of a convolutional distortion and as such, can be in part addressed by channel-oriented normalizations. However, in many instances, room impulse responses tend to have long tails compared to typical telephone channel responses – a fact reducing the effectiveness of normalizations operating on the level of short-term win-

*This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067.

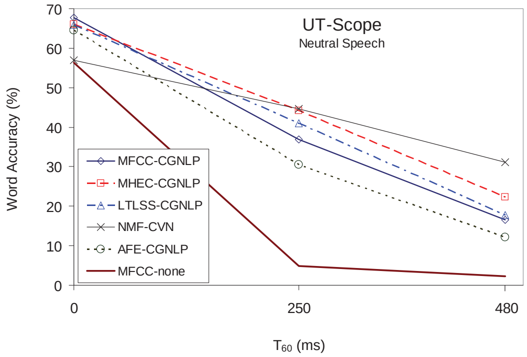

Figure 1: UT-Scope LVCSR; impact of reverberation on neutral speech recognition.

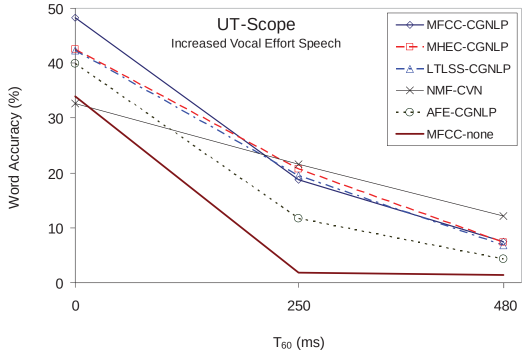

Figure 2: UT-Scope LVCSR; impact of reverberation on increased vocal effort speech recognition.

dows. The following feature normalizations are considered: *Distribution normalizations*: cepstral mean normalization (CMN), cepstral mean/variance normalization (CVN), Gaussianization (feature warping, *warp*) [14], histogram equalization (HEQ) [15], cepstral gain normalization (CGN) [16], and recently established quantile-based cepstral dynamics normalization (QCN) [17]. *Temporal filtering*: Relative spectral (RASTA) filtering [18] and recently proposed modified low-pass RASTA filtering (RASTALP) [9].

In particular, RASTA has been reported to have a potential to reduce the impact of reverberation on ASR [5]. In our previous study, RASTALP – a modified RASTA filter approximating the low-pass component of the original RASTA [9] and the high-pass portion by CMN or other segment-based normalizations [9, 19] was presented. Compared to the original high order RASTA filter, RASTALP requires significantly lower (2^{nd}) filter order, which results in considerable reduction of the transient effects typical for RASTA filtering. The combination of CMN–RASTALP outperformed RASTA in LVCSR on neutral and Lombard speech tasks in clean and noisy conditions [19].

4. Experimental Results

Two different speech corpora are utilized in this study – UT-Scope [20] and CLSD‘05 [21]. Both databases contain neutral and increased vocal effort speech. The increased vocal effort was induced by exposing the subjects to background noise, yielding a so called Lombard effect speech [22]. Lombard effect results in the increase of vocal effort and mean fundamental frequency, and affects also a number of other speech parameters [8, 21, 23, 24]. While the casue of the vocal effort increase is different for Lombard speech and speech produced in distant speaker-to-listener conditions, in both cases, the speech modifications result from the alteration of the auditory feedback. Due to the physiological mechanisms, the increased vocal effort goes hand in hand with changes of inherently related speech production parameters. Subglottal pressure and tension in the laryngeal musculature in higher vocal effort cause increase of mean fundamental frequency F_0 [25], which has been observed for altered auditory feedback both due to noise (Lombard effect) [8] and distant speaker-to-listener communication [4]. Increased vocal intensity is accompanied by the jaw lowering, which in turn causes an upward shift of the first formant F_1 [26]. Both migration of spectral energy and spectral center of gravity to higher frequencies [23], as well as flattening of the spectral tilt, are also typical for increased vocal effort in loud and Lombard speech [8]. Considering these similarities, clean Lombard speech seems to be a good approximation of the increased vocal effort speech observed in reverberation, and, hence, is used in this study.

4.1. UT-Scope Speech Corpus

The Lombard effect portion of the UT-Scope speech database contains neutral (modal) speech and speech produced with various levels of increased vocal effort [20]. The increased vocal effort was induced by playing three types of noises for subjects through headphones, and speech was captured by a close-talk microphone, yielding high signal-to-noise ratio (SNR) recordings. This allows for analysis of increased vocal effort speech with the inducing noise being excluded from the signal. The noise types used are: (i) highway car noise (speed 65 mph, windows half open) (ii) crowd noise, and (iii) pink noise. Highway and crowd noises were played at 70, 80, and 90 dB sound pressure level (SPL), pink noise at 65, 75, and 85 dB SPL. Sessions from 31 native speakers of American English (25 females, 6 males) are employed in the ASR experiments.

4.2. CLSD‘05 Speech Corpus

The Czech Lombard Speech Database (CLSD‘05) [21] comprises recordings of neutral speech and speech uttered in simulated noisy conditions (90 dB SPL of car noise). Similar as in UT-Scope, the noise samples were played through headphones, and speech was collected by a close-talk microphone. Sessions from 26 native speakers of Czech (12 females, 14 males) are utilized in the ASR experiments.

4.3. AIR Database

Two different reverberant test conditions are simulated by convolving the speech material with measured room impulse responses (RIR) from the Aachen Impulse Response (AIR) Database [27]. The RIR samples for a meeting room as well as an office are used with dimensions of $8.0 \times 5.0 \times 3.1$ m^3 and $5.0 \times 6.4 \times 2.9$ m^3, and reverberation times of T_{60}=250 ms and T_{60}=480 ms, respectively. Source-to-microphone distance is 2.8 m in the meeting room, and 3.0 m in the office.

4.4. UT-Scope LVCSR Experiments

Both UT-Scope and CLSD‘05 ASR systems utilize HTK for acoustic modeling, and the acoustic front-end features contain 13 static cepstral coefficients, including c_0, and their first and second order time derivatives.

A triphone recognizer with an SRILM trigram language model

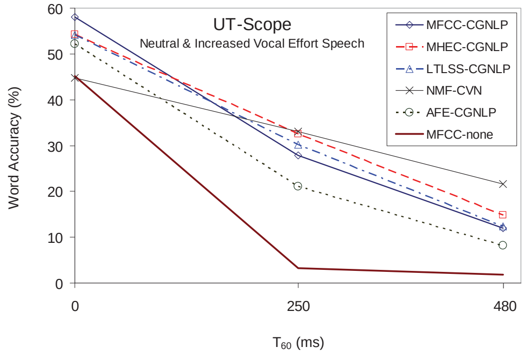

Figure 3: UT-Scope LVCSR; impact of reverberation on speech recognition on pooled neutral and increased vocal effort speech.

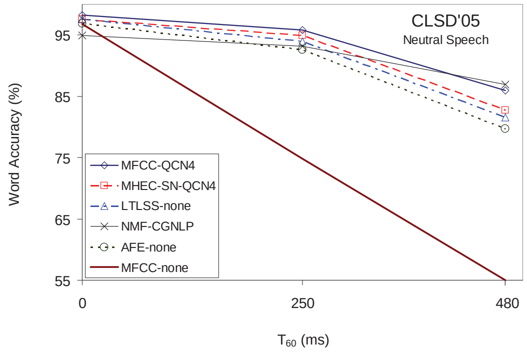

Figure 4: CLSD‘05 Digit Recognition; impact of reverberation on neutral speech recognition.

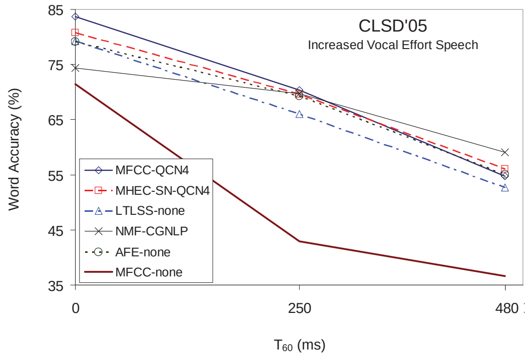

Figure 5: CLSD‘05 Digit Recognition; impact of reverberation on increased vocal effort speech recognition.

(LM) is trained on the TIMIT database. Here, 32-mixture TIMIT acoustic models are adapted towards UT-Scope with a combination of maximum likelihood linear regression (MLLR) adaptation, and maximum a posteriori (MAP) adaptation. The adaptation data is drawn from the UT-Scope *clean neutral samples*. Sessions of the adaptation set subjects are withdrawn from the evaluations. The test set contains sessions from 3 male and 19 female subjects. A total of 100 phonetically balanced TIMIT-like sentences produced in the neutral condition, and 20 TIMIT sentences produced in nine noise type/level conditions are available for each subject.

The ASR setups are evaluated on (i) *anechoic sets* – neutral speech and anechoic increased vocal effort speech produced in 70, 80, and 90 dB SPL of simulated highway and crowd noise, and 65, 75, and 85 dB of pink noise (the noise is not present in the recordings); (ii) the previous sets, reverberated with the RIR sample (T_{60} = 250 ms) from the AIR database; (iii) sets from *i* reverberated with the RIR sample (T_{60} = 480 ms) from the AIR database. This totals in 30 evaluation sets. The initial ASR system with MFCC–CVN front-end provides performance of 91.7 % word accuracy (*Acc*) on the anechoic neutral set. Since the focus of this study is on the effects of increased vocal effort and reverberation on acoustic modeling in ASR, the remainder of the paper reports word accuracies with LM being bypassed.

4.5. CLSD‘05 Small Vocabulary Experiments

A monophone recognizer is trained on the Czech SPEECON database [28]. The recognizer comprises 43 context-independent monophone models and two silence models. The models are trained on large vocabulary material from the Czech SPEECON database [28]. The task is to recognize 10 Czech digits (16 pronunciation variants) presented in connected digits utterances. The neutral test set comprises a total of 6353 words and the increased vocal effort test set 11663 words. Similar as in the UT-Scope case, the neutral and increased vocal effort sets are presented in the anechoic (i.e., original) and reverberant (T_{60} = 250 ms and T_{60} = 480 ms) conditions.

4.6. Results and Discussion

This section presents the observations made in the UT-Scope LVCSR and CLSD‘05 digit recognition experiments. Since the number of the ASR evaluation tasks, as well as the number of feature extraction strategies considered is extensive, in the following paragraphs we attempt to only summarize the overall trends and main outcomes of the experiments.

UT-Scope LVCSR Experiments: In the first step, efficiency of the normalizations from Sec. 3 included in an MFCC front-end was studied for anechoic and reverberated sets comprising pooled neutral and increased vocal effort samples. With increasing reverberation time T_{60}, the ASR performance severely deteriorates for all front-end setups. In all conditions, the combination of CMN and RASTALP (CMN-RASTA) outperformed traditional RASTA. The combinations CGN-RASTALP and QCN4-RASTALP consistently ranked among the top four normalizations in all scenarios, and ten out of twelve best performing front-ends utilized RASTALP filtering. Since CGN-RASTALP preceded, in the terms of word accuracy, QCN4-RASTALP in two out of three scenarios, it is considered to be the most efficient normalization in this evaluation. Detailed results of this experiment can be found in [12].

In the next step, selected feature extraction strategies were evaluated in combination with four normalizations (no normalization, CMN, CVN, and the best performing normalization identified in the previous paragraph – CGN-RASTALP). CMN and CVN are chosen to represent the common choice in many ASR engines. Front-ends mentioned in Sec. 1 and MHEC incorporating spectral subtraction (*MHEC-SS*), sub-band normalization (*MHEC-SN*), or both (*MHEC-SS-SN*) were combined with normalizations and evaluated on anechoic and reverberated sets. On *anechoic data sets*, once combined with any normalization, MFCC reached a superior performance. LTLSS, MHEC, and MHEC-SN ranked second behind MFCC. NMF provided inferior performance to all other front-ends. For *reverberated data* ($T_{60} = 250$ ms), MHEC-SS and MHEC-SN performed best, followed by NMF and MHEC. ETSI-AFE performed inferior to other front-ends. On *reverberated data* ($T_{60} = 480$ ms), NMF established highest accuracy, followed by the four MHEC configurations. CGN$_{LP}$ is most beneficial for all extraction strategies, except for NMF, which benefits most from CVN. Hence, NMF is paired with CVN and all other front-ends employ CGN$_{LP}$ in the subsequent analysis.

Third, feature extraction strategies are paired with their respective ‘optimal’ normalizations and evaluated separately for neutral, increased vocal effort, and pooled sets in anechoic, T_{60}=250 ms, and T_{60}=480 ms reverberation conditions. For comparison, performance of a baseline MFCC front-end without any normalization (denoted MFCC-none) is also evaluated. As can be seen in Fig. 1 and 2, MFCC-CGN$_{LP}$ establishes the best performance in anechoic conditions for both neutral and increased vocal effort speech while ranking fourth behind NMF, MHEC, LTLSS in $T_{60} = 250$ ms. On the other hand, NMF provides inferior performance in anechoic conditions (even dropping below the baseline performance for increased vocal effort speech) but matches the top-performing front-ends in $T_{60} = 480$ ms and clearly

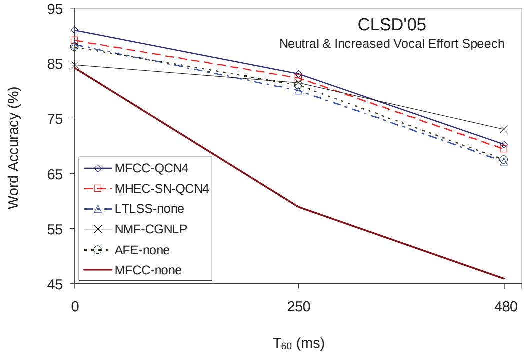

Figure 6: CLSD‘05 Digit Recognition; impact of reverberation on speech recognition on pooled neutral and increased vocal effort speech.

dominates in $T_{60} = 480$ ms. Fig. 3 suggests that the MHEC front-end would be the best choice for a recognizer operating in varying reverberation and vocal effort conditions, as MHEC-CGN$_{LP}$ provides the most balanced performance for pooled neutral and increased vocal effort speech in anechoic and reverberated conditions.

CLSD‘05 Small Vocabulary Experiments: Unlike in the case of UT-Scope, CLSD‘05 experiments focus on the small vocabulary digit recognition task. In addition, CLSD‘05 captures Czech spoken language and the recognizer utilizes monophone acoustic models. These differences allow for analysis of how transferable are the observations made in the previous paragraphs to another language and recognition task domain[1]. In the CLSD‘05 experiments, the front-end feature extraction strategies (MFCC, MHEC, LTLSS, NMF, AFE) were combined with all normalizations from Sec. 3 but feature warping and histogram equalization (those two provided a suboptimal performance in the initial experiments).

Figures 4, 5, and 6 depict the performance of each feature extraction strategy paired with the respective best performing normalization. It can be seen that the overall ASR performance is considerably higher here due to the simplicity of the task (digit recognition vs. LVCSR). Also, the recognition deterioration is much milder when switching from anechoic to T_{60}=250 ms conditions. Similarly, the small vocabulary recognition accuracy reduces less when switching from neutral to increased vocal effort speech. It can be assumed that the word models in general small vocabulary task are more easily distinguishable in the acoustic feature space and are less affected by the speech deterioration due to reverberation compared to the LVCSR triphone acoustic models.

Surprisingly, all feature extraction strategies in the CLSD‘05 task paired with different ‘optimal’ normalizations than in the UT-Scope task. Similar to UT-Scope, MFCC maintains the best performance on anechoic neutral and increased vocal effort sets. This transfers also to the T_{60}=250 ms condition here. Similar to UT-Scope, NMF dominates in $T_{60} = 480$ ms for both types of speech. On CLSD‘05, the combination of MFCC and QCN represents the best choice across most of the conditions, MHEC-SN and QCN being the second best front-end.

It can be seen that while plain MFCC front-end does not deal well with either reverberation or increased vocal effort in both UT-Scope and CLSD‘05, it becomes quite competitive when paired with a well chosen normalization – see Fig. 2 in the UT-Scope task and Fig. 4–6 in the CLSD‘05 task.

5. Conclusion

This study analyzed the individual as well as combined impacts of reverberation and increased vocal effort on large and small vocabulary recognition in English and Czech languages, respectively. Robustness of several standard and state-of-the-art feature extraction techniques and normalizations was evaluated in the varying reverberation/vocal effort conditions. Several similar trends were observed for both the English large and Czech small vocabulary tasks. In particular, it was observed that ASR performance deteriorated with increasing vocal effort and reverberation time. However, this deterioration is milder in the low vocabulary task, presumably due to the lower confusability of the small vocabulary models in the acoustic feature space. When combined with an ‘optimal’ normalization, MFCC front-end always outperformed other schemes in anechoic conditions, both for neutral and increased vocal effort speech. In addition, its performance remained competitive with other front-ends in T_{60}=250 ms for increased vocal effort speech in the large vocabulary task as well as both neutral and increased vocal effort speech in the small vocabulary task. NMF front-end was consistently inferior in anechoic conditions, but became competitive in T_{60}=250 ms and dominated in $T_{60} = 480$ ms in all evaluations. Recently established MHEC feature extraction front-end provided well balanced performance in both LVCSR and small vocabulary tasks and state-of-the-art QCN and CGN-RASTALP normalizations ranked among the top choices for most front-ends considered in this study.

[1] In an ideal world, the previous best front-end configuration might be expected to provide a sustained superior performance also here.

6. REFERENCES

[1] A. K. Nabelek, T. R. Letowski, and F. M. Tucker, “Reverberant overlap- and self-masking in consonant identification,” *J. Acoust. Soc. Am.*, vol. 86, pp. 1259–1265, Oct. 1989.

[2] Q. Lin, C. Che, D.-S. Yuk, L. Jin, B. deVries, J. Pearson, and J. Flanagan, “Robust distant-talking speech recognition,” in *Proc. IEEE ICASSP*, vol. 1, May 1996, pp. 21–24.

[3] D. Gelbart and N. Morgan, “Double the trouble: handling noise and reverberation in far-field automatic speech recognition,” in *Proc. ICSLP*, Sept. 2002, pp. 2185–2188.

[4] D. Pelegrín-García, B. Smits, J. Brunskog, and C.-H. Jeong, “Vocal effort with changing talker-to-listener distance in different acoustic environments,” *J. Acoust. Soc. Am.*, vol. 129, no. 4, pp. 1981–1990, Apr. 2011.

[5] B. Kingsbury and N. Morgan, “Recognizing reverberant speech with RASTA-PLP,” in *Proc. IEEE ICASSP*, vol. 2, Apr. 1997, pp. 1259–1262.

[6] K. Lebart, J. Boucher, and P. Denbigh, “A new method based on spectral subtraction for speech dereverberation,” *Acta Acustica*, vol. 87, pp. 359–366, 2001.

[7] K. Kumar, R. S. B. Raj, and R. M. Stern, “Gammatone sub-band magnitude-domain dereverberation for ASR,” in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[8] J. H. L. Hansen, “Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition,” *Speech Commun.*, vol. 20, no. 1-2, pp. 151–173, Nov. 1996.

[9] H. Bořil and J. H. L. Hansen, “UT-Scope: Towards LVCSR under Lombard effect induced by varying types and levels of noisy background,” in *Proc. IEEE ICASSP*, Prague, Czech, May 2011, pp. 4472–4475.

[10] S. O. Sadjadi and J. H. L. Hansen, “Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions,” in *Proc. IEEE ICASSP*, May 2011, pp. 5448–5451.

[11] “Speech processing, transmission and quality aspects (stq), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm,” in *ETSI standard document-ETSI ES 202 050 v1.1.1*, 2002.

[12] O. Sadjadi, H. Bořil, and J. H. L. Hansen, “A comparison of front-end compensation strategies for robust LVCSR under room reverberation and increased vocal effort,” to appear in *Proc. IEEE ICASSP*, Mar. 2012.

[13] M. Wu and D. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Trans. ASLP*, vol. 14, pp. 774–784, May 2006.

[14] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification,” in *Proc. A Speaker Odyssey - The Speaker Recognition Workshop*, Jun. 2001, pp. 213–218.

[15] S. Dharanipragada and M. Padmanabha, “A nonlinear unsupervised adaptation technique for speech recognition,” in *Proc. ICSLP*, Oct. 2000, pp. 556–559.

[16] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, “Cepstral gain normalization for noise robust speech recognition,” in *Proc. IEEE ICASSP*, vol. 1, May 2004, pp. 209–212.

[17] H. Bořil and J. H. L. Hansen, “Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments,” *IEEE Trans. ASLP*, vol. 18, no. 6, pp. 1379–1393, Aug. 2010.

[18] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Trans. SAP*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[19] H. Bořil, F. Grézl, and J. H. L. Hansen, “Front-end compensation methods for LVCSR under Lombard effect,” in *Proc. INTERSPEECH*, 2011.

[20] J. H. L. 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. ASLP*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[21] H. Bořil, “Robust speech recognition: Analysis and equalization of Lombard effect in Czech corpora,” Ph.D. dissertation, CTU in Prague, Czech Rep., http://www.utdallas.edu/∼hynek, 2008.

[22] J.-C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Am.*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[23] Y. Lu and M. Cooke, “Speech production modifications produced by competing talkers, babble and stationary noise,” *J. Acoust. Soc. Am.*, vol. 124, no. 5, pp. 3261–3275, Nov. 2008.

[24] M. Garnier, “Communication in noisy environments: From adaptation to vocal straining,” Ph.D. dissertation, Univ. of Paris VI, France, 2007.

[25] R. Schulman, “Dynamic and perceptual constraints of loud speech,” *J. Acoust. Soc. Am.*, vol. 78, no. S1, pp. S37–S37, 1985.

[26] ——, “Articulatory dynamics of loud and normal speech,” *J. Acoust. Soc. Am.*, vol. 85, no. 1, pp. 295–312, Jan. 1989.

[27] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Proc. IEEE DSP*, Jul. 2009, pp. 1–5.

[28] D. Iskra, B. Grosskopf, K. Marasek, H. van den Huevel, F. Diehl, and A. Kiessling, “SPEECON – Speech databases for consumer devices: Database specification and validation,” in *Proc. of LREC‘2002*, 2002, pp. 329–333.