A COMPARISON OF FRONT-END COMPENSATION STRATEGIES FOR ROBUST LVCSR UNDER ROOM REVERBERATION AND INCREASED VOCAL EFFORT

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

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA

{sadjadi, hynek, john.hansen}@utdallas.edu

ABSTRACT

Automatic speech recognition is known to deteriorate in the presence of room reverberation and variation of vocal effort in speakers. This study considers robustness of several state-of-the-art front-end feature extraction and normalization strategies to these sources of speech signal variability in the context of large vocabulary continuous speech recognition (LVCSR). A speech database recorded in an anechoic room, capturing modal speech and speech produced at different levels of vocal effort, is reverberated using measured room impulse responses and utilized in the evaluations. It is shown that the combination of recently introduced mean Hilbert envelope coefficients (MHEC) and a normalization strategy combining cepstral gain normalization and modified RASTA filtering (CGN_RASTALP) provides considerable recognition performance gains for reverberant modal and high vocal effort speech.

Index Terms— Feature normalization, robust acoustic features, robust speech recognition, room reverberation, vocal effort

1. INTRODUCTION

Room reverberation poses various detrimental effects on spectro-temporal characteristics of speech signals, among which self- and overlap-masking are most notable [1]. In a reverberant enclosure, sound waves arrive at the receiver (e.g., ears or microphone) via a direct path, and via multiple paths and directions after reflecting off walls and objects defining the acoustic enclosure. The reflections arriving within $50 - 80$ ms after the direct sound are called early reflections, which tend to build up to a level louder than the direct sound and cause an internal smearing effect known as the self-masking effect. The echoes reaching the receiver after the early reflections are called late reflections, which tend to smear the direct sound over time and mask succeeding sounds. This phenomenon is commonly referred to as the overlap-masking effect, and has been shown to be the primary cause of degraded speech recognition performance in both human listeners [1] and automatic speech recognizers [2, 3].

Not only can reverberation cause signal distortion, it also results 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 deal with both signal distortions introduced by reverberation itself and also the signal variability due to the increased vocal effort, which is induced by reverberation. Several studies have considered individual effects of room reverberation [2, 3, 5–7] and increased vocal effort [8, 9] on ASR, and reported compensation strategies for their alleviation. However, to the best of our knowledge, this study is the first to consider the individual as well as the combined effects of reverberation and increased vocal effort on ASR. Robustness of various conventional and recently proposed feature extraction/compensation techniques are evaluated in the context of large vocabulary continuous speech recognition (LVCSR) 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 LVCSR experiments. The feature extraction schemes are paired with a number of popular cepstral normalizations and also recently proposed RASTALP temporal filtering. It is noted that this represents the first attempt to evaluate MHEC in an ASR task and analyze robustness of RASTALP to reverberation.

It is shown that post-processing the MHEC with cepstral gain normalization (CGN) [12] combined with modified low-pass RASTA filtering [9], which has been recently introduced for robust ASR under noisy *Lombard effect* conditions, results in considerable improvement in performance under reverberant modal and high vocal effort speech.

2. MEAN HILBERT ENVELOPE COEFFICIENTS: MHEC

MHEC features (Fig. 1) have been shown to be an effective alternative to MFCC for robust SID under reverberant mismatched conditions [10]. The fourth and fifth stages (in the dashed box) in Fig. 1 are optional and employed to suppress the reverberation self and overlap-masking effects.

First, the pre-emphasized reverberant speech signal $r(t)$ is decomposed into 26 bands through a 26-channel Gammatone filterbank. Next, the Hilbert envelope $e_r(t, j)$ 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 $e_r(t, j)$ is blocked into frames of 25 ms duration with a skip rate of 10 ms. To estimate the temporal envelope amplitude in frame m, the sample mean $R(m, j)$ is computed. Note that $R(m, j)$ is a measure of the spectral energy at the center frequency of the j^{th} channel, and therefore provides a short-term spectral representation of the speech signal $r(t)$. Next, in each channel, the envelope trajectories are normalized using the long-term average computed over the entire utterance, yielding $R_n(m, j)$. This stage 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 due to early reflections has been suppressed. The overlap-masking effect, which is the long-term effect of reverberation 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 speech spectrum $\hat{S}(m, j)$. The last stage (i.e., logarithm and DCT) is commonly used in extraction of conventional cepstral features such as MFCCs. 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

The following feature normalizations considered in this study are typically applied in cepstral or log spectral domain in an effort to reduce the impact of speaker, channel, and environmental noise mismatch on speech

This project was funded by AFRL through a subcontract to RADC Inc. under FA8750-09-C-0067. Approved for public release; distribution unlimited.

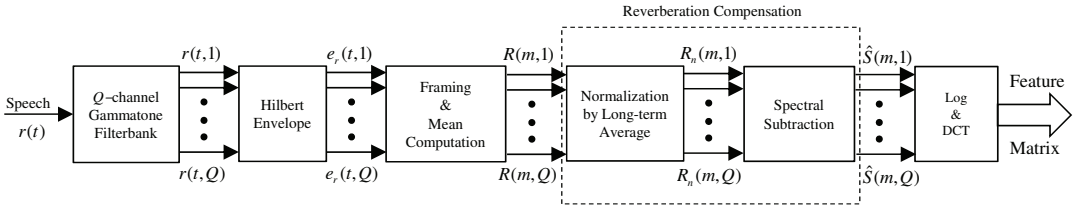

Fig. 1. Block diagram of the MHEC feature extraction framework. The symbols represent the output signals at each stage.

systems.

Distribution normalizations:

- Moment normalizations: cepstral mean normalization (CMN), cepstral mean/variance normalization (CVN), Gaussianization (feature warping, warp) [14], histogram equalization (HEQ) [15],
- Range normalizations: cepstral gain normalization (CGN) [12], quantile-based cepstral dynamics normalization (QCN) [16].

Temporal filtering:

- Relative spectral (RASTA) filtering [17],
- Modified low-pass RASTA filtering (RASTALP) [9].

Due to the deconvolution properties of the log spectral/cepstral domain, signal distortions caused by changes in environmental acoustics, microphone/channel path, as well as speech production changes can be modeled as a variation in feature distribution means and variances. In this sense, the distribution normalizations operating in this domain have potential to reduce also the impact of reverberation, characterized by the room impulse response (RIR), and increased vocal effort, reflected in speech intensity and spectral slope.

RASTA band-pass temporal filtering suppresses speech signal components that are assumed to vary either too slowly or quickly to be attributed to speech. RASTA has also potential to reduce the impact of reverberation [5] by smoothing the additional feature envelope peaks due to the signal reflection from the room walls.

In our recent study, RASTALP – a modified RASTA filter that approximates the low-pass portion of the original RASTA by a smoothing low pass filter [9] and the high-pass portion by CMN or other segment-based normalizations [9, 18] was introduced. Compared to the original high order band-pass RASTA filter, RASTALP is a filter of significantly lower (2^{nd}) order, which helps reduce transient effects typical for RASTA filtering. The combination of CMN–RASTALP was shown to outperform RASTA in LVCSR on neutral and high vocal effort tasks presented in clean and noisy conditions [18].

4. EXPERIMENTAL RESULTS

4.1. Speech Corpus

The test samples are drawn from the Lombard effect portion of the UT-Scope speech database that contains neutral (modal) speech and speech produced with various levels of increased vocal effort [19]. Lombard effect (LE) represents a phenomenon where speakers adjust their speech production in order to maintain intelligible communication in noisy environments [20]. It is reflected in the increase of vocal effort, mean fundamental frequency, and a number of other speech parameters [8]. While the cause behind the vocal effort increase is different for Lombard speech and speech produced in reverberant distant speaker-to-listener conditions, the impact on speech production parameters is, due to the physiological mechanisms, in many aspects similar. Increased subglottal pressure and tension in the laryngeal musculature in higher vocal effort cause increase of mean fundamental frequency F_0 [21], which has been observed for both Lombard speech [8] and distant speaker–listener speech [4]. Increased vocal effort is typically accompanied by the jaw lowering, which results in the upward shift of the first formant F_1 in frequency [22]. Both migration of spectral energy and spectral center of gravity to higher frequencies, as well as flattening of the spectral tilt, are also typical for increased vocal effort in loud and Lombard speech [8].

Three types of noisy backgrounds were played to subjects through headphones in an ASHA certified anechoic sound booth to induce increased vocal effort: (i) highway car noise (speed 65 mph, windows half open) (ii) crowd noise, and (iii) pink noise. Highway and crowd noises were produced through headphones 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 US English (25 females, 6 males) are used in the ASR experiments. Each session comprises 100 phonetically balanced read sentences from the TIMIT database produced by each subject in the neutral condition, and 20 TIMIT sentences produced in each of the nine noise type/level conditions.

To simulate different reverberant conditions, RIR samples extracted from the Aachen Impulse Response (AIR) Database [23] are convolved with the test material. Two RIR's with distinct source-to-microphone distances (d_{SM}) are used including meeting and office rooms. More information about the RIR's is summarized in Table 1. Here, the DRR denotes the direct-to-reverberant ratio which is dependent on the distance between the sound source and microphone.

4.2. Experimental Setup

A triphone recognizer utilizing HTK acoustic models and the SRILM trigram language model (LM) is trained on the TIMIT database. The feature vector is formed using 13 static cepstral coefficients, including c_0, and their first and second order time derivatives. To alleviate the acoustic/channel mismatch between TIMIT and UT-Scope, the 32-mixture triphone models are adapted towards UT-Scope using combined maximum likelihood linear (MLLR) adaptation and maximum a posteriori (MAP) adaptation on a subset of *clean neutral speech* UT-Scope recordings. Adaptation set subjects are excluded from evaluations. The test sets contain sessions from 3 male and 19 female subjects.

The ASR setups are evaluated on (i) *anechoic sets* – neutral speech and anechoic Lombard speech produced in 70, 80, and 90 dB SPL of simulated highway and crowd noise, and 65, 75, and 85 dB of pink noise (noise was produced through headphones and does not appear in the LE recordings); (ii) set (i) reverberated with the first RIR sample from the AIR database (see the first row in Table 1) with T_{60} = 250 ms; (iii) sets from *i* reverberated with the second RIR sample from the AIR database (see the second row in Table 1) with T_{60} = 480 ms. This yields a total of 30 evaluation sets. The initial ASR system utilizing MFCC–CVN front-end establishes performance on the anechoic neutral set at 91.7 % word accuracy (*Acc*). Since our focus is on comparing the efficiency of the front-end strategies in the context of acoustic modeling, the remainder of the paper reports word accuracies obtained from the acoustic model decoding with the LM bypassed.

4.3. Results and Discussion

In this section, efficiency of selected feature extraction strategies combined with feature normalizations discussed in Sec. 3 is evaluated on the anechoic and reverberated neutral/increased vocal effort speech sets. Since the evaluation of all possible front-end combinations in all con-

Table 1. Properties of two selected RIR samples from AIR database.

Room Type	Dimension (m^3)	d_{SM} (m)	T_{60} (s)	DRR (dB)
Meeting	$8.0 \times 5.0 \times 3.1$	2.80	0.25	2.89
Office	$5.0 \times 6.4 \times 2.9$	3.0	0.48	-0.89

Table 2. Comparison of cepstral compensations in MFCC front-end; sorted by performance in descending order.

Rank-Ordered *MFCC-Norm* Setups; Accuracy (%)					
Anechoic		T_{60} = 250 ms		T_{60} = 480 ms	
CGN_{LP}	58.0	**$QCN4_{LP}$**	28.2	$warp_{LP}$	12.1
QCN4	57.8	HEQ_{LP}	28.2	HEQ_{LP}	12.1
$QCN4_{LP}$	57.3	**CGN_{LP}**	27.8	**CGN_{LP}**	12.0
CMN_{LP}	57.0	QCN4	27.7	**$QCN4_{LP}$**	12.0
CVN_{LP}	56.8	$warp_{LP}$	27.1	warp	11.4
CVN	56.3	CVN_{LP}	26.2	CVN	11.3
CMN	55.6	CVN	26.1	HEQ	11.3
$warp_{LP}$	55.4	warp	25.7	CVN_{LP}	11.3
HEQ_{LP}	55.3	CMN_{LP}	25.4	QCN4	11.2
HEQ	54.7	HEQ	25.1	CMN_{LP}	10.9
warp	53.7	CMN	24.8	CMN	10.6
CGN	52.3	RASTA	21.6	RASTA	8.5
RASTA	50.4	CGN	19.5	CGN	5.6
none	45.1	none	3.3	none	1.8

ditions would yield an extensive number of results, the experiments are broken down into three stages. First, a common feature extraction strategy (MFCC) is paired with all available normalizations and evaluated on anechoic and reverberated sets (mixture of neutral/increased vocal effort samples). Second, selected feature extraction strategies are evaluated in four setups (no normalization, CMN, CVN, and the best normalization found in the first stage). CMN and CVN normalizations are chosen to represent the common choice in many ASR engines. Third, feature extraction strategies paired with respective best performing normalizations are evaluated in detail separately for neutral and increased vocal effort sets in anechoic, T_{60}=250 ms, and T_{60}=480 ms reverberation conditions.

Normalizations in MFCC front-end: performance of raw MFCC system and MFCC combined with normalizations discussed in Sec. 3 is summarized in Table 2 for anechoic and reverberated sets (neutral and increased vocal effort samples are pooled together and given equal weight in the overall word accuracy *Acc*). Setups denoted *Norm_RASTALP* or $Norm_{LP}$ represent a combination of the normalization *Norm* followed by RASTALP. *QCN4* denotes a QCN setup where 4^{th} and 96^{th} quantiles represent the dynamic range to be normalized [16]. Table 2 displays front-end configurations sorted by their performance for the anechoic and the two reverberant conditions. It can be seen that with increasing reverberation time T_{60}, the ASR performance severely deteriorates for all setups. In all conditions, CMN_RASTALP outperforms traditional RASTA. CGN_RASTALP and QCN4_RASTALP consistently rank among the top four normalizations in all scenarios, and ten out of twelve top front-ends utilize RASTALP filtering. Since CGN_RASTALP precedes QCN_RASTALP in two out of three scenarios, it is selected to accompany CMN and CVN in the subsequent evaluations.

Comparison of feature extraction strategies: front-ends mentioned in the introduction and MHEC incorporating spectral subtraction (*MHEC-SS*), MHEC with sub-band normalization (*MHEC-SN*), and MHEC combining both SS and SN (*MHEC-SS-SN*) are paired with selected normalizations and evaluated on anechoic and reverberated sets in Fig. 2, 3, and 4. On *anechoic data sets*, once combined with any normalization, raw MFCC reaches a superior performance. LTLSS, MHEC, and MHEC-SN rank second behind MFCC. NMF provides inferior performance to all other front-ends. On *reverberated data* (T_{60} = 250 ms), MHEC-SS and MHEC-SN perform best, followed by NMF and MHEC. The fact that MHEC-SS outperforms MHEC-SN for both reverberation times supports the fact that the reverberation overlap-masking poses more deleterious impact on ASR performance than the reverberation self-masking [1, 6]. ETSI-AFE ranks last among the setups. On *reverberated data* (T_{60} = 480 ms), NMF establishes highest Acc, followed by the four MHEC setups. When averaging the performance of individual front-ends across the anechoic and reverberant conditions, CGN_{LP} is most beneficial for all extraction strategies, except for NMF, which benefits most from CVN. Hence, in the subsequent performance analysis, NMF is paired with CVN and all other front-ends utilize CGN_{LP}.

Neutral versus increased vocal effort speech: Following the intuition, increasing reverberation time results in steep performance degradation on

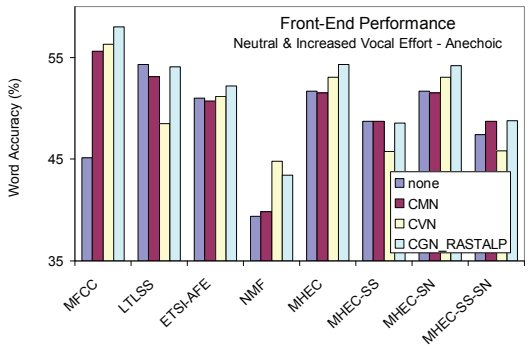

Fig. 2. Comparison of front-end strategies combined with normalizations; *anechoic* neutral & increased vocal effort sets.

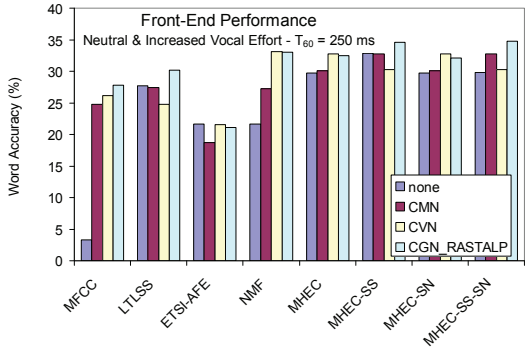

Fig. 3. Comparison of front-end strategies combined with normalizations; *reverberated* neutral & increased vocal effort sets; (T_{60} = 250 ms).

both neutral (Fig. 5) and increased vocal effort (Fig. 6) speech. Presence of increased vocal effort further deteriorates a neutral-trained ASR. For both neutral and increased vocal effort, MFCC paired with CGN_{LP} provides highest accuracy (67.7%) on anechoic data, followed by MHEC-SN and MHEC (66.1%), and LTLSS (65.9%). NMF reaches the lowest accuracy (56.9%). In reverberation of T_{60} = 250 ms, the top ranking front-ends on neutral speech are MHEC-SS-SN (48.0%), MHEC-SS (47.8%), followed by NMF (44.6%) and MHEC (44.2%). For increased vocal effort, the top competitors are similar, with MHEC-SS-SN and NMF (21.6%), MHEC-SS (21.5%), and MHEC (20.8%). In both cases, ETSI-AFE provides lowest performance (lagging by over 10% on neutral and 15% on increased vocal effort behind MFCC). In reverberation of T_{60} = 480 ms, NMF reaches superior accuracy on both neutral and increased vocal effort speech (31.1% and 12.1%), followed by MHEC-SS (26.0% and 9.1%) and MHEC-SS-SN (25.9% and 9.0%). As in the previous case, ETSI-AFE provides the lowest performance of all systems.

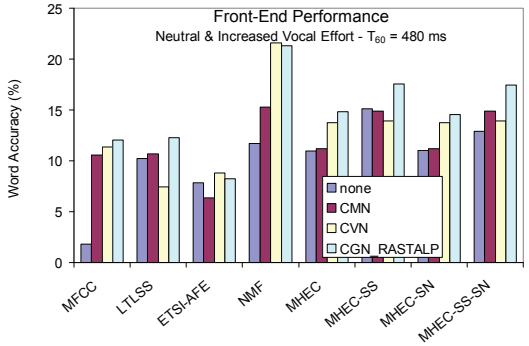

Fig. 4. Comparison of front-end strategies combined with normalizations; *reverberated* neutral & increased vocal effort sets; (T_{60} = 480 ms).

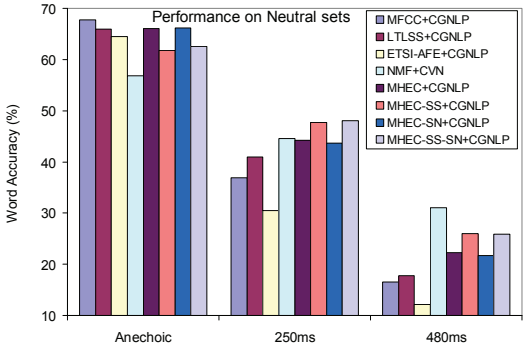

Fig. 5. Accuracy as a function of T_{60}: front-end strategies combined with CGN_RASTALP or CVN normalization; neutral sets.

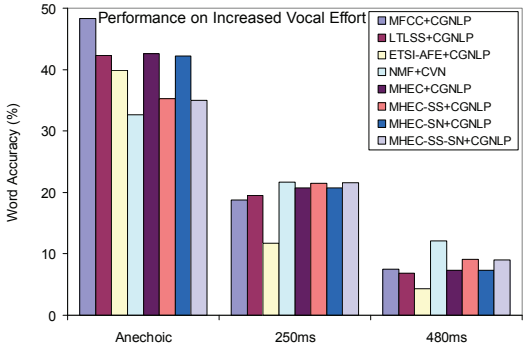

Fig. 6. Accuracy as a function of T_{60}: front-end strategies combined with CGN_RASTALP or CVN normalization; increased vocal effort sets.

The experimental results can be summarized as follows. RASTALP filtering consistently benefits ASR performance for neutral and increased vocal effort speech in anechoic/reverberate conditions, appearing in ten out of twelve most efficient normalizations paired with MFCC (Table 2). Combination CGN-RASTALP outperforms CMN and CVN in seven out of eight front-ends (ranking second behind CVN in NMF). MHEC-SN, MHEC, and LTLSS reach the second best performance on anechoic data, following MFCC. MHEC-based front-ends dominate in $T_{60} = 250$ ms, being closely followed by NMF. NMF performs best in the highest reverberation time ($T_{60} = 480$ ms), followed by the MHEC setups. NMF's impressive performance in the last condition is spoiled by its lagging by more than 10% Acc behind MFCC in anechoic neutral conditions. LTLSS provides comparable performance to MHEC setups on anechoic data, but loses in reverberated conditions. ETSI-AFE lags behind MFCC, MHEC and LTSS on anechoic data and is consistently worst on both neutral and increased vocal effort reverberated data. This is not surprising given that ETSI-AFE was designed with the focus on ASR under additive noise conditions.

5. CONCLUSION

This study analyzed individual and joint impact of reverberation and increased vocal effort on automatic speech recognition. Robustness of several traditional and state-of-the-art feature extraction techniques and normalizations was evaluated on neutral (modal) speech and speech produced with various levels of increased vocal effort. Speech samples were reverberated by measured room impulse responses with two different reverberation times. Recently proposed mean Hilbert envelope coefficients (MHEC) and CGN_RASTALP normalization were, in conjunction, shown to outperform state-of-the-art long-term log spectral subtraction (LTLSS) MFCC and ETSI advanced front-end (ESI-AFE) cepstra in all reverberant conditions and both speech modalities. MHEC provided better or comparable performance to nonnegative matrix factorization (NMF) in reverberated conditions ($T_{60} = 250$ ms) on neutral and increased vocal effort speech. NMF provided superior performance in strong reverberation ($T_{60} = 480$ ms), yet, at the same time, failed in anechoic conditions (lagging by over 10% on neutral and 15% absolute word accuracy on increased vocal effort behind MFCC). The results suggest that the proposed combination of MHEC and CGN_RASTALP provides a balanced contribution to recognition performance across various reverberation and vocal effort conditions and has a good potential to benefit a broad scope of ASR applications.

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] 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.

[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] 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.

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

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

[19] 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.

[20] 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.

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

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

[23] 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.