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

The use of Locally Normalized Cepstral Coefficients (LNCC) to improve speaker recognition accuracy in highly reverberant rooms

Víctor Poblete [1], Juan Pablo Escudero [1], Josué Fredes[2], José Novoa[2], Richard M. Stern[3], Simon King [4], Néstor Becerra Yoma [2]

[1] Institute of Acoustics, Universidad Austral de Chile, Valdivia, Chile
[2] Speech Processing and Transmission Laboratory, Electrical Engineering Department, University of Chile, Santiago, Chile
[3] Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, USA
[4] Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

vpoblete@uach.cl, juan.escudero@alumnos.uach.cl, jfredes@ing.uchile.cl, jose.novoa@ug.uchile.cl, rms@cs.cmu.edu, Simon.King@ed.ac.uk, nbecerra@ing.uchile.cl

Abstract

We describe the ability of LNCC features (Locally Normalized Cepstral Coefficients) to improve speaker recognition accuracy in highly reverberant environments. We used a realistic test environment, in which we changed the number and nature of reflective surfaces in the room, creating four increasingly reverberant times from approximately 1 to 9 seconds. In this room, we re-recorded reverberated versions of the Yoho speaker verification corpus. The recordings were made using four speaker-to-microphone distances, from 0.32m to 2.56m. Experimental results for a speaker verification task suggest that LNCC features are an attractive alternative to MFCC features under such reverberant conditions, as they were observed to improve verification accuracy compared to baseline MFCC features in all cases where the reverberation time exceeded 1 second or with a greater speaker-microphone distance (i.e. 2.56 m).

Index Terms: LNCC, reverberation, distant speaker microphone, speaker verification.

1. Introduction

For many state-of-the-art algorithms, the quality of acoustic conditions for speech communication in enclosed spaces including reverberation is important [1-2]. Inside such spaces, considered as a transmission channel, a time varying speech signal emitted from a speaker may reach a listener through several paths. In general the speech reaching the listener is not an ideal copy of the original speech [3]. Propagated sound is inevitably reflected by nearby surfaces [4] resulting in higher-order reflections [5]. This superposition is perceived not as individual echoes, but as a single acoustic entity, described as *reverberation* [6-7]. The auditory system distinguishes between direct sound, early lateral reflections and late reverberation [8,9]. Direct sound is heard within an interval of 25-35ms. The early reflections [10-11] have arrival times of less than 50ms [12-13]. Such reflections increase the effective SNR and improve the intelligibility of speech (the amount of an utterance that is understood) [14-15]. Late reverberation components, with arrival times greater than 50-100ms, degrade speech intelligibility [10]. The integration of the direct sound, early reflections, and late reverberation, is accomplished in the brain [16,17]. Reverberant distortion increases with the speaker-to-listener distance, r [18]. The direct sound level decreases by 6 dB for every doubling of r [18], and D/R, the ratio of energy in the direct to reverberant field, decreases by 6 dB for every doubling of distance:

$$D/R = -6\log_2\left(\frac{r}{r_c}\right) = 20\log\left(\frac{r_c}{r}\right) \quad dB \tag{1}$$

where r_c is the critical distance of the room, at which D/R equals 0 dB (direct and reverberant energies are equal) [19]. The value of r_c is calculated as in [19-20]:

$$r_c = 0.06\left(\frac{G\cdot V}{RT}\right)^{\frac{1}{2}} \quad m \tag{2}$$

where V and RT are the volume (m^3) and reverberation time (s), respectively; and G is the directivity factor of the source. The directivity index DI is measured in a hemi-anechoic room, and its value is obtained according to [21]. Thus, by using:

$$DI(f) = 10\log_{10}(G(f)) \tag{3}$$

G is calculated for the frequencies 500, 1kHz, and 2kHz [5,22], using the equation [23]:

$$G(f) = 10^{DI(f)/10} \tag{4}$$

Following [24], the decay rate τ of the impulse response curve of a listening space is linearly proportional to RT in seconds [25]:

$$RT = 6.91\,\tau \tag{5}$$

Theoretically, the decay curve is calculated using the root-mean-squared (rms) pressure in a time window from 0 to T and is given by [4]:

$$p_{rms} = \sqrt{\frac{1}{T}\int_0^T p^2\,dt} \tag{6}$$

The STI index [26], a predictor of intelligibility, varies in the range from 0 (bad) to 1 (excellent) and is calculated following [27]. Speech de-reverberation algorithms are classified in four categories: signal-based, feature-based, model-based, and decoder-based [2]. Feature-based approaches include RASTA filtering [28] and Cepstral Mean Normalization (CMN) [29].

http://dx.doi.org/10.21437/Interspeech.2016-1277

In this paper, our main motivation is to compare LNCC with MFCC combined CMN or RASTA processing. We are not attempting to exhaustively compare all known techniques to address reverberation. Most research on speech recognition and verification in reverberant environments has been carried out by modifying existing databases using simulation, and the effect of the speaker-microphone distance has typically been neglected. The purpose of the current paper is to assess the influence of reverberant distortion on LNCC [30] including the effect of speaker-microphone distance. The use of LNCC is motivated by the robustness observed for normal-hearing listeners in adapting to listening spaces with moderate amounts of reverberation [7,22]. In the frequency domain (Fig. 1), a local normalization is performed by dividing the outputs of two filters: (1) the numerator filter is triangular, and essentially the same as that used to derive MFCC features [31]; (2) the denominator filter captures energy at adjacent frequencies. They are defined by the equations (7) and (8):

$$Num_i(f)=\begin{cases} -\frac{2}{B}\left|f-f_i^C\right|+1\,, \text{ when } \left|f-f_i^C\right|\le\frac{B}{2} \\ 0 \quad , \text{otherwise} \end{cases} \tag{7}$$

$$Den_i(f)=\begin{cases} \frac{2}{B}(1-d_{min})\left|f-f_i^C\right|+d_{min}\,, \text{ when } \left|f-f_i^C\right|\le\frac{B}{2} \\ 0 \quad , \text{otherwise} \end{cases} \tag{8}$$

For any auditory channel i, the locally normalized channel energy LN_i is achieved by dividing:

$$LN_i=\frac{Num\_Energy_i}{Den\_Energy_i} \tag{9}$$

where Num_Energy_i and Den_Energy_i represent the energy captured by the filters in Eq. (7) and Eq. (8). LNCC features have already been evaluated for robustness with respect to spectral tilt in transmission channels and additive noise [30, 31]. Figure 2 plots spectral envelopes of speech recorded in Room 1 at a distance 2.56 m from the sound source, estimated by the LN filterbank, and compares this to the corresponding response of the conventional Mel-scale filterbank typically used to derive MFCCs [32].

2. Experimental procedures

2.1. Characteristics of the reverberation chamber

We constructed several reverberant environments and re-recorded sequentially with a single microphone, different versions of the Yoho speech corpus [33]. The room volume and its total interior surface are 203 m^3 and 215 m^2, respectively. The longest diagonal distance is 11 m, as seen in Fig. 3. Four different reverberant fields were created by varying the number and nature of absorbent objects and surfaces in a single room. We label these virtual rooms as Room 1, Room 2, Room 3 and Room 4: highly reverberant, reverberant, moderately reverberant, and mildly reverberant, respectively. The RT values of the chamber, as a function of frequency, were measured according to [34] (see Fig. 4). Four distances between the playback loudspeaker (Bose V-201) and microphone (Shure PG-81) r are used: 0.32, 0.64, 1.28 and 2.56 m. In each r, RT is measured with a sound level meter (Cesva SC310) and software (Cesva Capture Studio). The averaged values of RT vary from 0.9, 2.0, 3.0 to 9.3s for Rooms 4, 3, 2 and 1, respectively (averaged values at 500, 1000, and 2000Hz bands [5,22]). The acoustic indices selected to quantify the intelligibility are: RT, SNR, STI, and D/R, as listed in Table 1.

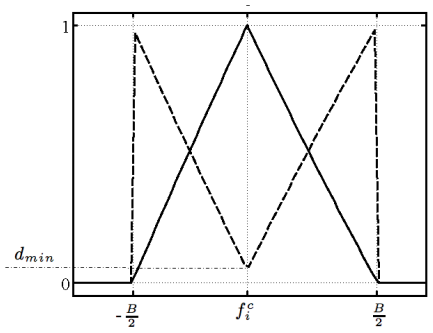

Figure 1: *Frequency responses of numerator (solid line) and denominator (dashed line) filters.*

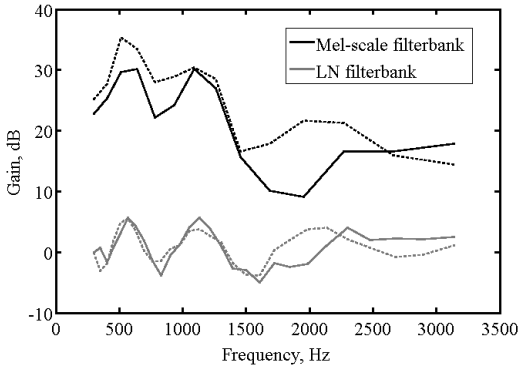

Figure 2: *Spectral envelopes for a single frame of clean voiced speech (solid line) using Mel-scale filterbank (upper figure), and LN filterbank (lower figure). In solid lines, responses without reverberation; dashed lines, with reverberation.*

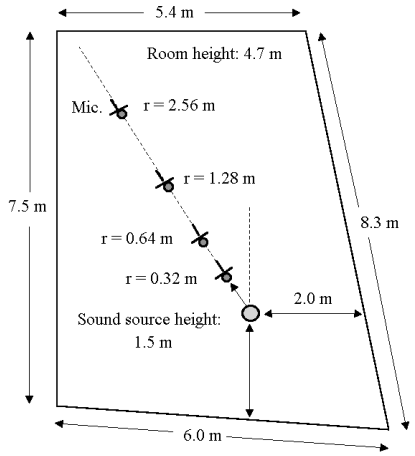

Figure 3: *Schematic for the reverberation chamber.*

2.1.1. Schematic for the experimental setup

The experimental setup is illustrated in Fig. 3. The room height is an average height above the source because the ceiling surface is an inclined plane. Fig. 5 is a calculated example of Eq. (6) obtained from room impulse responses measured in this reverberation room. For example, as is seen

in Fig. 5, the slope of the decay curve for Room 1 describes a reverberation time much longer than that of Room 4.

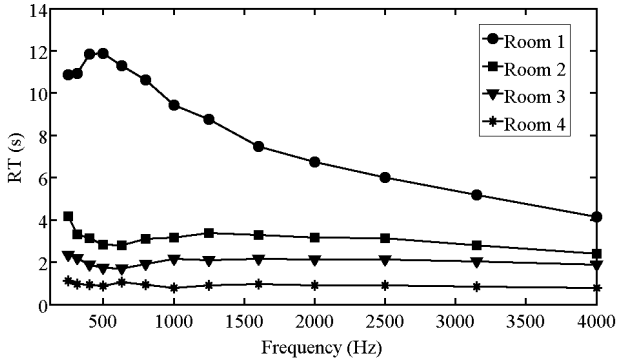

Figure 4: *Averaged values of the reverberation times as a function of frequency*.

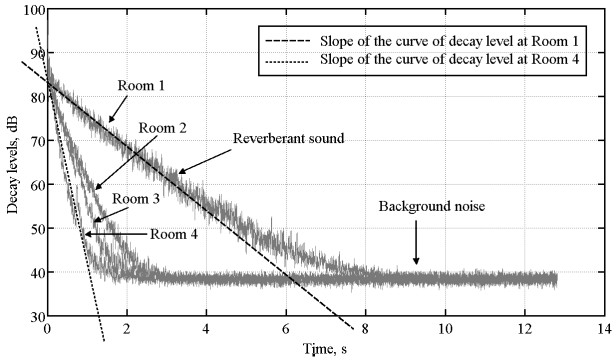

Figure 5: *Curves of decay level for each room*.

2.1.2. Speaker verification experiments

LNCC performance under reverberant distortion was tested using a text-independent speaker verification paradigm. All experiments used the entire Yoho speech corpus. Features were extracted using LNCC and MFCC processing. The frame duration in all cases was 25 ms with a 50% overlap. A frequency range from 200 to 3860 Hz was covered by 14 triangular filters uniformly arranged on a Bark scale, in the case of MFCCs, and in the case of the LNCC features they are computed using 28 pairs of numerator and denominator filters uniformly arranged on a Bark scale, with dmin = 0.001, and B = 3.5 Barks. The baseline system for clean speech produces 0.56% and 0.71% EER with MFCC and LNCC features, respectively. These experiments were carried out using the ALIZE library and LIA-SpkDet toolkit [35]. This software is based on a classical Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system [36]. The Universal Background Model (UBM) is trained using background impostor speakers, with 256 Gaussian components using diagonal covariance matrices. A speaker-dependent Gaussian Mixture Model (GMM) is generated for each speaker by employing maximum *a posteriori* (MAP) adaptation [36].

3. Speaker verification results

3.1. Room 1: Highly reverberant

Figure 6 shows Equal Error Rates (EERs) obtained using LNCC features or standard MFCC features, in highly reverberant conditions. Results for MFCC or LNCC each in combination with CMN or RASTA are also shown.

LNCC features provide relative reductions in EER as high as 31.1%, 25.4%, 20.6%, and 7.2%, compared to MFCC, at speaker-microphone distances 0.32, 0.64, 1.28, and 2.56m, respectively. These results suggest that LNCC is more robust than MFCC to the highly reverberant condition. When CMN is applied to MFCC or LNCC, there is no reduction in EER, compared to MFCC or LNCC alone. When RASTA is applied to LNCC or MFCC, instead of CMN, relative reductions in EER are as high as 33.2% at a speaker-microphone distance 0.64 m for LNCC+RASTA over MFCC+RASTA. We observe that LNCC benefits from this additional normalization as much as MFCC does.

Table 1. Reverberation and intelligibility conditions.

Room	r *(m)*	r_c *(m)*	RT (s)	SNR (dB)	STI	D/R (dB)
1	0.32	0.62	9.35	4.2	0.54	5.8
	0.64	0.63	9.24	1.4	0.44	-0.2
	1.28	0.63	9.28	0.4	0.34	-6.2
	2.56	0.62	9.43	-0.2	0.29	-12.3
2	0.32	1.09	3.05	7	0.66	10.7
	0.64	1.09	3.10	3.5	0.60	4.6
	1.28	1.09	3.08	1	0.51	-1.4
	2.56	1.09	3.07	-1	0.41	-7.4
3	0.32	1.35	2.00	9.7	0.77	12.5
	0.64	1.33	2.06	5.0	0.7	6.4
	1.28	1.36	1.98	1.2	0.59	0.5
	2.56	1.34	2.04	-0.7	0.53	-5.6
4	0.32	2.01	0.85	14.1	0.88	16.2
	0.64	1.93	0.98	8.7	0.85	9.6
	1.28	1.94	0.97	4.0	0.79	3.6
	2.56	1.90	1.01	0.3	0.69	-2.6

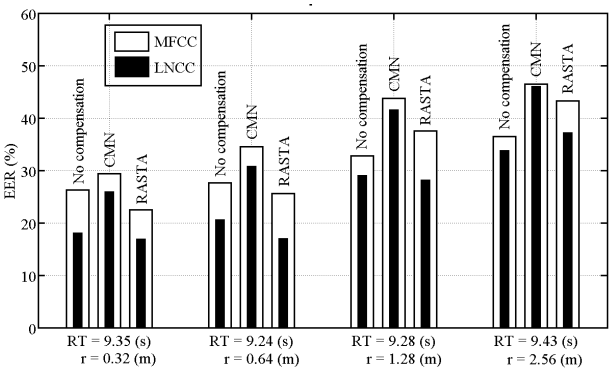

Figure 6: *Comparison of performance in Room 1.*

3.2. Room 2: Reverberant

Figure 7 summarizes the results in Room 2. Again, LNCC provides relative reductions in EER, over MFCC, this time of 36.2%, 32.1%, 19.4%, and 13.28% at speaker-microphone distances 0.32, 0.64, 1.28, and 2.56m. CMN is not effective in combination with MFCC or LNCC. LNCC+RASTA outperforms MFCC, MFCC+CMN, and MFCC + RASTA.

3.3. Room 3: Moderately reverberant

Figure 8 presents EERs in Room 3. LNCC provides relative reductions in EER over MFCC of 28.5%, 19.2%, 18.9%, and

19.8%, for speaker-microphone distances 0.32, 0.64, 1.28, and 2.56m. At 0.32m, LNCC (+CMN or +RASTA) is no longer superior to MFCC (+CMN or +RASTA).

3.4. Room 4: Mildly reverberant

Figure 9 summarizes the performance of LNCC features under mild reverberation. At 0.32 and 0.64 m, LNCC (+CMN or +RASTA) is no longer superior to MFCC (+CMN or +RASTA). Nevertheless, LNCC still provides relative reductions over standard MFCC features of between 5.6% and 22.9%, for larger speaker-microphone distances (1.28 and 2.56m). For the milder conditions of reverberation (at closer speaker-microphone distances) LNCC does not improve speaker recognition accuracy. These results are replotted as a function of *D/R* in Fig. 10.

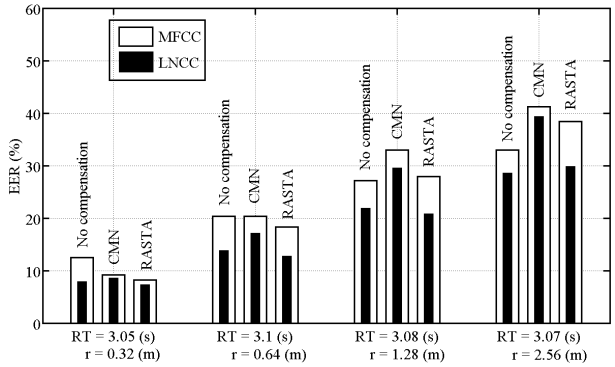

Figure 7: *Comparison of performance in Room 2.*

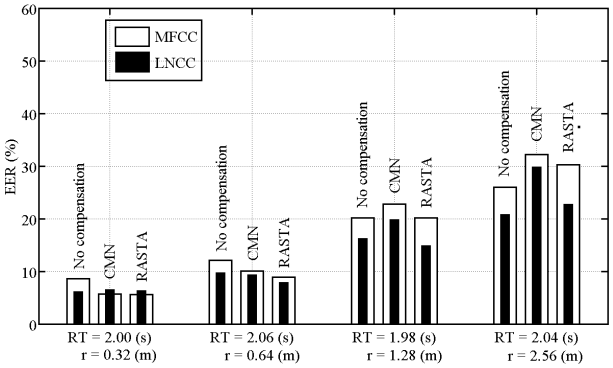

Figure 8: *Comparison of performance in Room 3.*

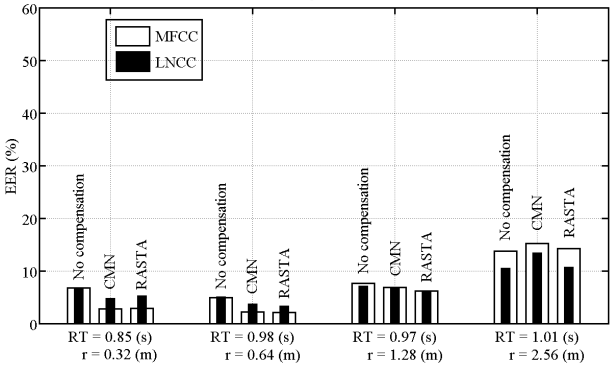

Figure 9: *Comparison of performance in Room 4.*

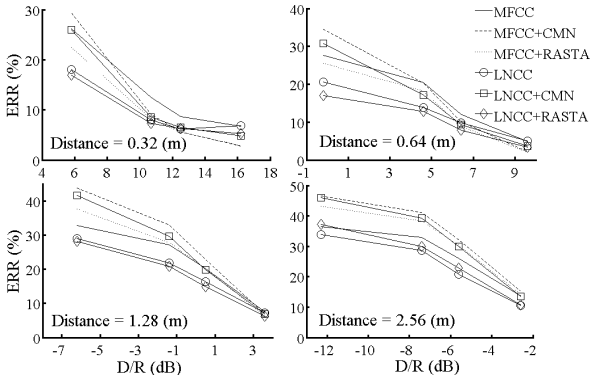

Figure 10: *EER (%) versus D/R (dB).*

4. Discussion

By manipulating the reverberation time of a room, and re-recording data at several speaker-microphone distances, we have been able to measure the impact of reverberation typical of listening spaces commonly used for spoken communication. In general, we found that LNCC features produce lower EERs than MFCC features (each in combination with either CMN or RASTA) in a speaker verification task, under most conditions considered, with the exception of those conditions with the least reverberation and closest speaker-microphone distances. The use of LNCC features alone, at 2.56 m, provides relative reductions in EER, over standard MFCC of 22.9%, 19.8, 13.3%, and 7.2%, in Rooms 4, 3, 2 and 1, respectively.

5. Conclusions

Our speaker-verification results using the Yoho database demonstrate that LNCC is more robust than MFCC for all speaker-microphone distances in Rooms 1, 2 and 3. We conclude that for the majority of speaker-microphone distances, LNCCs provide better performance than MFCCs. Observations on the extent to which CMN and RASTA improve accuracy when added to LNCC processing are mixed, although in general RASTA is helpful and CMN is not. We also note that, independent of reverberation time, in all four reverberant Rooms, at the largest speaker-microphone distance of 2.56 m (where the indexes SNR, STI, as well as D/R correspond to the poorest intelligibility for speech), LNCCs alone with no compensation are more robust than standard MFCCs. We conclude that LNCC features can be an attractive alternative to MFCC, which can also be applied in other tasks of pattern recognition where occurs reverberant distortions, poor intelligibility, or when the speaker-microphone distance is varying. LNCCs appear to be particularly attractive features for distant speech processing in a variety of real, reverberant environments.

6. Acknowledgements

The research reported here was partly funded by grant DID-UACh 2015-63, and Conicyt projects PIA ACT 1120 and Fondecyt 1151306. This work was partially supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

7. References

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, T., and W. Kellermann, “Making machines understand us in reverberant rooms,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 114-126, 2012.

[2] K. Kinoshita, M. Delcroix, S. Gannot, E.A.P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," EURASIP Journal on Advances in Signal Processing, vol. 7, pp. 1-19, 2016.

[3] T. Houtgast and H.J.M. Steeneken, "A Review of the MTF concept in room acoustics and its use for estimating speech-intelligibility in auditoria," Journal of the Acoustical Society of America, vol. 77, no. 3, pp. 1069-77, 1985.

[4] H. Kuttruff, "A simple iteration scheme for the computation of decay constants in enclosure with diffusely reflecting boundaries," Journal of the Acoustical Society of America, vol. 98, no. 1, pp. 288-293, 1995.

[5] I. Arweiler and J.M. Buchholz, "The influence of spectral characteristics of early reflections on speech intelligibility," Journal of the Acoustical Society of America, vol. 130, no. 2, pp. 996-1005, 2011.

[6] M. Sayles and I.M. Winter, "Reverberation challenges the temporal representation of the pitch of complex sounds," Neuron, vol. 58, no. 5, pp. 789-801, 2008.

[7] S. Devore, A. Ihlefeld, K. Hancock, B. Shinn-Cunningham, and B. Delgutte, "Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain," Neuron, vol. 62, no. 1, pp. 123-134, 2009.

[8] J.S. Bradley and H. Sato, "On the importance of early reflections for speech in rooms," Journal of the Acoustical Society of America, vol. 113, no. 6, pp. 3233-44, 2003.

[9] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," IEEE Transactions on the Audio Speech and Language Processing, vol. 17, no. 4, pp. 1-12, 2009.

[10] T. Hidaka, Y. Yamada, and T. Nakagawa, "A new definition of boundary point between early reflections and late reverberation in room impulse responses," Journal of the Acoustical Society of America, vol. 122, no. 1, pp. 326-332, 2007.

[11] J. Catic, S. Santurette, and T. Dau, "The role of reverberation-related binaural cues in the externalization of speech," Journal of the Acoustical Society of America, vol. 138, no. 2, pp. 1154-67, 2015.

[12] J.S. Bradley and G.A. Soulodre, "Objective measures of listener envelopment," Journal of the Acoustical Society of America, vol. 98, no. 5, part 1, pp. 2590-97, 1995.

[13] J.F. Li, R.S. Xia, Q. Fang, A.J. Li, J.L. Pan, and Y.H. Yan, "Effect of the division between early and late reflections on intelligibility of ideal binary-masked speech," Journal of the Acoustical Society of America, vol. 137, no. 5, pp. 2801-10, 2015.

[14] H. Sato, J.S. Bradley, and M. Morimoto, "Using listening difficulty ratings of conditions for speech communication in rooms," Journal of the Acoustical Society of America, vol. 117, no. 3, part 1, pp. 1157-67, 2005.

[15] A. Lindau, L. Kosanke, and S. Weinzierl, "Perceptual evaluation of model and signal based predictors of the mixing time in binaural room impulse responses," Journal of the Audio Engineering Society, vol. 60, no. 1, pp. 887-898, 2012.

[16] M.M. Murray and L. Spierer, "Multisensory integration: What you see is where you hear," Current Biology, vol. 21, no. 6, pp. R229-R231, 2011.

[17] S. Gordon-Salant and P.J. Fitzgibbons, "Profile of auditory temporal processing in older listeners," Journal of Speech Language and Hearing Research, vol. 42, no. 2, pp. 300-311, 1999.

[18] E. Larsen, N. Iyer, C.R. Lansing, and A.S. Feng, "On the minimum audible difference in direct-to-reverberant energy ratio," Journal of the Acoustical Society of America, vol. 124, no. 1, pp. 450-61, 2008.

[19] A.W. Bronkhorst and T. Houtgast, "Auditory distance perception in rooms," Nature, vol. 397, no. 6719, pp. 517-520, 1999.

[20] A.W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," Acustica, vol. 86, no. 1, pp. 117-128, 2000.

[21] International Standard ISO 3745:2012 Part E. Acoustics. Determination of sound power levels and sound energy levels of noise sources using sound pressure. Precision methods for anechoic rooms and hemi-anechoic rooms.

[22] A.K. Nabelek and T.R. Letowski, "Similarities of vowels in nonreverberant and reverberant fields," Journal of the Acoustical Society of America, vol. 83, no. 5, pp. 1891-99, 1988.

[23] D.A Bies and C.H. Hansen, *Engineering Noise Control: Theory and Practice*. Spon Press, fourth edition, London.

[24] R. Ratnam, D.L. Jones, B.C. Wheeler, W.D. O’Brien, Jr., C. R. Lansing, and A.S. Feng, "Blind estimation of reverberation time," Journal of the Acoustical Society of America, vol. 114, no. 5, pp. 2877–92, 2003.

[25] R. Ratnam, D.L. Jones, and W.D. O’Brien, Jr., “Fast algorithms for blind estimation of reverberation time,” IEEE Signal Processing Letters, vol. 11, no. 6, pp. 537-540, 2004.

[26] K.L. Payton, R.M. Uchanskbi, L.D. Braida, “Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing,” Journal of the Acoustical Society of America, vol. 95, no. 3, pp. 1581-92, 1994.

[27] H. J. Steeneken and T. Houtgast, “A physical method for measuring speech-transmission quality,” Journal of the Acoustical Society of America, vol. 67, pp. 318-326, 1980.

[28] H. Hermansky and N. Morgan, “RASTA processing of speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 2, no. 4, pp. pp.578-589, 1994.

[29] Liu, F., R.M. Stern, X. Huang, and A. Acero, “Efficient cepstral normalization for robust speech recognition,” Proceedings of ARPA Human Language Technology Workshop, 1993.

[30] V. Poblete, F. Espic, S. King, R. M. Stern, F. Huenupán, J. Fredes, and N. Becerra Yoma, "A perceptually-motivated low-complexity instantaneous linear channel normalization technique applied to speaker verification," Computer Speech and Language, vol. 31, pp. 1-27, 2015.

[31] J. Fredes, J. Novoa, V. Poblete, S. King, R.M. Stern, and N. Becerra Yoma, “Robustness to additive noise of Locally Normalized Cepstral Coefficients in speaker verification,” Proceedings Interspeech 2015, pp. 3011-3015, Dresden.

[32] S. B. Davis and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.

[33] J.P. Campbell and A. Higgings, YOHO Speaker Verification. Linguistic Data Consortium, Philadelphia, PA, 1994.

[34] International Standard ISO 3382:2008. Acoustics. Measurement of the reverberation time of rooms with reference to other acoustical parameters.

[35] J.F. Bonastre, et al., “ALIZE/spkdet: A state of the art open source software for speaker recognition,” Odyssey, 2008.

[36] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19–41, 2000.