Objective evaluation of static beamforming on the quality of speech in noise

Vincent Grimaldi, Gilles Courtois and Hervé Lissek
Swiss Federal Institute of Technology (EPFL), Signal Processing Laboratory (LTS2), Lausanne, Switzerland.

Eleftheria Georganti
Sonova AG, Stäfa, Switzerland.

Summary
Beamforming is commonly used in devices such as in hearing instruments (HI) with a view to reducing noise. It consists in focusing on audio signals arising from a prescribed steering direction while suppressing those coming from other directions. The efficiency of beamformers on speech intelligibility has been demonstrated in various studies. Nevertheless, little is known about the effect on speech quality. This paper aims to assess the impact of static beamforming on the signal statistics and on speech quality in the context of hearing aids using four objective measures. Three of them are associated to temporal fine structure (TFS), spectral content evolution and envelope information, respectively. The well-known Hearing-Aid Speech Quality Index (HASQI) is used to assess speech distortion additionally. These assessments have been performed in various acoustic conditions in order to evaluate the impact of the room, the signal-to-noise ratio (SNR), and the direction of arrival of the speaker's voice. The measures indicate that the resort to beamforming tends to improve listening quality in addition to the speech intelligibility enhancement it provides.

PACS no. xx.xx.Nn, xx.xx.Nn

1. Introduction

In many communication applications such as in hearing instruments (HI), speech intelligibility can be impaired due to the presence of background noise. Consequently, various techniques are frequently used to improve the signal-to-noise ratio (SNR) with the aim to recover a better understanding on speech. Beamforming is a common technique of spatial filtering used for enhancing speech coming from a prescribed direction while eliminating noise coming from other directions [1]. A microphone array is used to take advantage of different spatial, temporal and frequency cues to create a beam at a desired direction. Beamforming approaches can be static, with a beam in a predetermined fixed direction, or adaptive by steering the beam in a desired direction [2, 3].

Improvements in terms of noise reduction and speech intelligibility were demonstrated in several studies [4, 5, 6], and for the particular case of hearing aids [7, 8]. Nevertheless little is known about the effect of beamforming on speech quality in noise. A non-intrusive method specifically designed for dereverberated signals was proposed in [9] aiming at assessing both speech intelligibility and quality. Speech distortion caused by a noise reduction algorithm and the additional effect of the beamformer was discussed in [10] showing that the use of a beamformer reduces speech distortion caused by a noise reduction algorithm. The distortion of various advanced beamforming techniques was also discussed in [11], revealing that the fixed beamformer produced the least distortion.

A review of various existing objective measures for speech enhancement by noise suppression algorithms were evaluated in [12] from which the Perceptual Evaluation of Speech Quality (PESQ) [13], log-likelihood ratio (LLR) [14] and frequency-weighted segmental SNR (fwSNRseg) [15] stood out. The Perceptual Objective Listening Quality Assessment (POLQA) was introduced [16] and shown to outperform PESQ, especially for wideband signals. In [17], the Hearing-Aid Speech Quality Index (HASQI) [18] and the PErception MOdel Hearing Impairment Quality (PEMO-Q-HI) [19] were demonstrated to outperform other measures for the quality assessment in the context of speech enhancement in hearing-aids.

In this work, the evaluation is focused on the effect of static beamforming on the speech quality in noisy environments in the context of HIs. Three metrics are introduced in order to investigate properties of recorded speech signals in diffuse babble noise and assess the quality in comparison to a clean reference. The HASQI is also included to measure speech distortion.

2. Static beamformer

In the following sections, we consider the particular case of hearing aids equipped with two omni-directional microphones per device, denoted front and back microphone respectively, as shown of Figure 1.

Figure 1: Hearing aid (Phonak Bolero^{TM} Q) equipped with two microphones

The static beamformer is implemented in both left and right HIs, as depicted on Figure 2. H_F2B is the relative transfer function from the front to the back microphone for a specific null direction set to 180°, in order to obtain a beam steering toward the front. H_F2B was obtained from head-related transfer functions (HRTF) measurements conducted in an anechoic chamber with the two HIs.

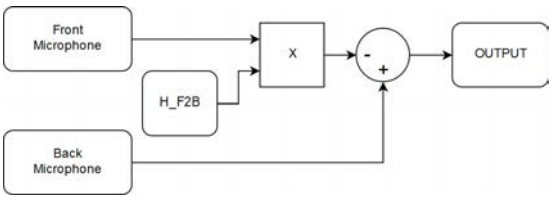

Figure 2: Static beamformer implementation for one HI

The beamformer as described in this paper is implemented as the conventional delay-and-sum processing, which allows to obtain a cardioid directivity pattern facing forward. Figure 3 depicts the directivity pattern in the left HI obtained from measurements at various frequencies with the implementation described on Figure 2. One should note that the maximum is obtained around 30°, which is due to the head effect. The effect is more pronounced for higher frequencies and a symmetric pattern could be observed at the right HI (not reported here). Without loss of generality, only the left HI is considered in this paper.

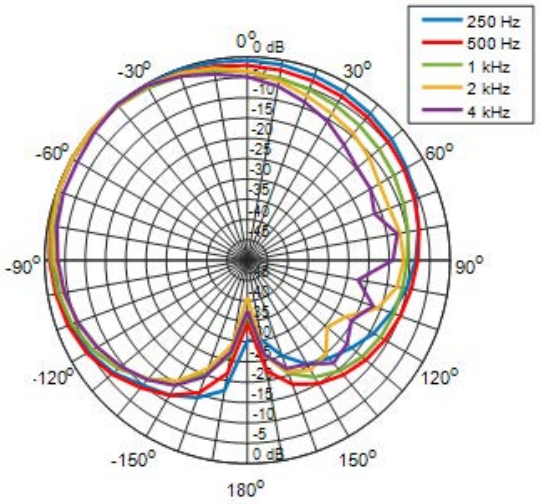

Figure 3: Directivity pattern of the static beamformer of the left HI with null set at azimuth 180° for various frequencies

In order to compensate for the high-pass effect resulting from the directional processing [20], a compensation filter is derived by measuring the magnitude responses of the beamformed and omni-directional signals to a broadband source at 0°, in a similar manner as in [10]. Nevertheless, full compensation results in an undesired boost of low-frequency noise. Consequently, low-frequency gains have been empirically reduced to obtain a trade-off between transparency and noise level.

3. Objective metrics

Various metrics for evaluating listening quality in noise are described in this section. They cover important properties of speech signals: the temporal fine structure (TFS), the frequency content and the envelope.

3.1. Temporal kurtosis

The kurtosis is a statistical measure that evaluates the shape of a probability distribution. The computed value is related to the tails of the distribution. A low value is associated with a more spread distribution. In this context, it is directly computed on the signals of interest in the time domain to assess the distributions of the amplitude of the samples. Therefore, this measure is an indicator related to the TFS. A clean and high quality speech signal should exhibit a narrow distribution. Conversely, a speech signal including a larger amount of noise is associated with a more flat distribution due to the presence of many additional intermediate amplitude values from the noise. Examples are depicted on Figure 4 for a speech signal recorded in a quiet (green) and a noisy (red) environments.

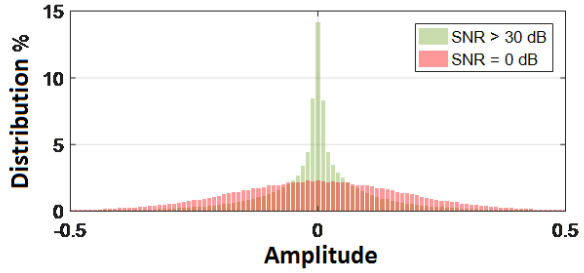

Figure 4: Distribution of the samples values for a clean speech (SNR $>$ 30 dB, green) and the same speech in noise (SNR $=$ 0 dB, red)

3.2. Variance of energy in the Bark bands

The evolution of the spectral content over time is another information to observe when evaluating speech quality. For this purpose, the variance of the energy in the Bark bands [21] is considered. Clean speech material exhibits higher variances due to the variations of the formant energy over time. The presence of stationary noise tends to reduce the value of the variance due to its slow varying statistics over time. Examples of energy variance in Bark bands are shown on Figure 5 for a clean (green) and a noisy (red) signal in a similar configuration. Note that this metric should not be used in the presence of a highly non-stationary masker, such as a competitive speaker.

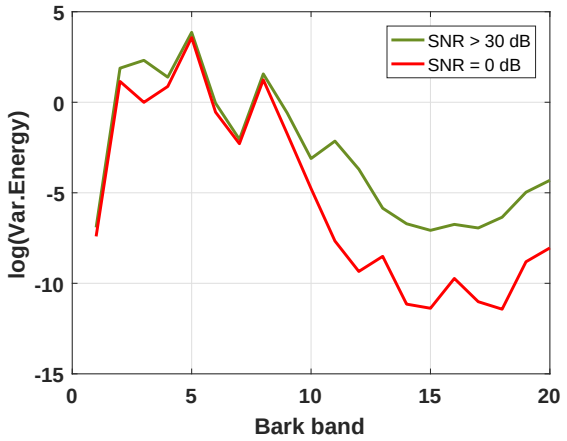

Figure 5: Variance of energy in the Bark bands for a clean speech (SNR $>$ 30 dB, green) and the same speech in noise (SNR $=$ 0 dB, red)

3.3. Spectral Centroid of the envelope

The envelope is another component of the signal that can be used to evaluate speech quality. More precisely, the spectrum of the envelope, called the modulation transfer function (MTF) [22], can be computed. The modulation information is usually located below 25Hz for speech content, and a peak around 4Hz corresponding to the syllabic rate can be noticed for clean speech [23]. The spectral centroid of the MTF is the frequency that represents the barycenter of the entire spectrum. A clean signal results in a low spectral centroid, whereas a noisier signal is associated with a shift toward higher frequencies due to the additional faster envelope modulations included in the noise. Examples are depicted on Figure 6 for a clean (green) and a noisy (red) signal in a similar configuration.

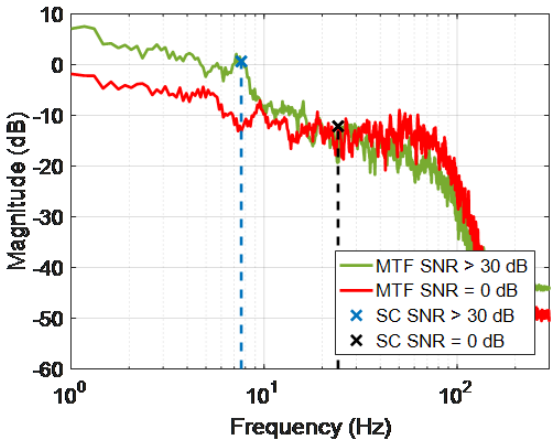

Figure 6: Spectral centroid (SC) of the MTF for a clean speech (SNR $>$ 30 dB, green) and the same speech in noise (SNR $=$ 0 dB, red)

3.4. HASQI

The HASQI [18] is an index specifically designed to evaluate speech quality in hearing aids. It is based on an auditory model that allows to include changes due to hearing loss in the evaluation. The measure is a combination of two nonlinear indices: cepstral correlation and correlation of the vibration of the modeled basilar membrane, and two linear indices: differences in spectra and differences in spectra slopes. The HASQI is then obtained by an empirical combination of those indices. A score of 1 means that no distortion is caused by the processing, while values decreasing toward 0 quantify the amount of distortion.

4. Method

The dataset used for the evaluation was obtained from measurements. This section describes the recording setup as well as the methods for computing the metrics.

4.1. Dataset

The setup used for the measurements is depicted on Figure 7. The stimulus consists of a 14-second male speech (sample rate $=$ 22050 Hz), played through a

HATS B&K type 4128 manikin (speaker) at a distance of two meters from a KEMAR manikin (listener) equipped with a pair of Phonak Bolero^{TM} Q hearing aids. Recordings of the speech source at various azimuths (0°, 30°, 60°, 90°and 120°) were conducted to assess the effect of the beamformer in the horizontal plane.

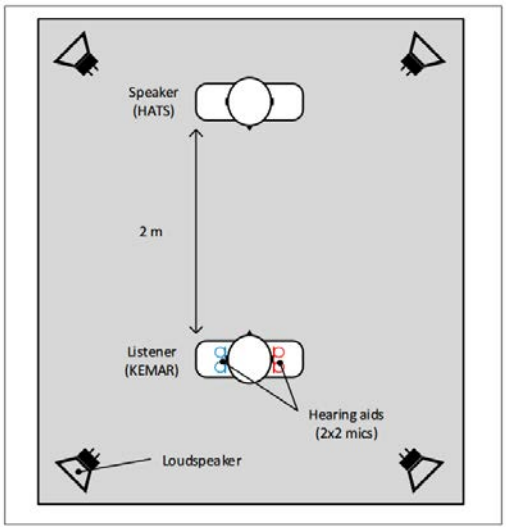

Figure 7: Setup of the recordings used for the dataset

Signals from the front and back omni-directional microphones of the left hearing aid are considered and allow to compute the beamformed signal in the frontal direction for the left HI. Two environments, a listening room (volume = 125 m^3, RT_{60} = 0.17 s at the listener position) and a classroom (volume = 262 m^3, RT_{60} = 0.59 s at the listener position) as well as two diffuse noise levels (SNR = 10 dB and SNR = 0 dB) are also considered. The noise consists of two different diffuse babbles for each of the environment. They are generated by playing babble recordings from four loudspeakers (Tannoy Reveal Active studio monitors) located at the corners of each room.

4.2. Reference and metrics

The clean omni-directional recording (SNR > 30 dB) in the front left microphone at a distance of two meters is used as a reference signal for each room. The objective metrics are computed for this reference signal as well as for the beamformed and unprocessed noisy signals for both environments and SNRs. The three first reported measures cover important properties of speech, namely the TFS, the spectral content and the envelope. The HASQI is used to evaluate the speech distortion, as a well-known reference index for speech quality for hearing aid applications.

5. Results

The measures for the temporal kurtosis, variance of energy in the Bark bands, spectral centroid of the MTF and HASQI are presented and discussed in this section.

5.1. Scores

The scores obtained from the objective metrics of interest are shown on Figure 8. The results are displayed as the distance of the score of the signal of interest to the score of the clean version. This is to evaluate the tendency of the processing to shift the properties closer or further from the reference. A lower bar is associated with more resemblance to the reference and consequently of a higher speech quality regarding the corresponding property.

5.2. Observations

The beamformer shifts the kurtosis to greater values for speech coming from angles up to 60° in the listening room and up to 90° in the classroom at both SNRs. The effect is more pronounced at the highest SNR. This is an indication that beamforming tends to improve the TFS resemblance to the reference for this range of angles in the look direction. At larger azimuths the kurtosis indicates a poorer quality regarding the TFS as could be expected.

Up to 90° the variance in the Bark bands is also increased to values closer to the clean reference statistics for both rooms and SNRs, meaning that this metric is improved by the use of a static beamformer. One should note that the babble used as masker in both environments is relatively stationary which is a condition for this metric to be relevant. Indeed, as mentioned previously, this measure would not be significant for a highly non-stationary noise and very low SNR.

The spectral centroid of the MTF is shifted down for azimuths up to 90° in the listening room and 30° in the classroom, indicating an improvement of the envelope characteristic in the steering direction with the use of beamforming.

Finally, as discussed in [10], the HASQI indicates that speech distortion is reduced by the beamformer, for azimuths up to 120° in the listening room and up to 30° in the classroom. In accordance with [11], improvements of quality owing to the beamformer are more pronounced in the less reverberant environment (listening room). This could relate to the fact improvements on speech intelligibility with beamforming are more moderate in reverberant environments [7].

Generally, the beamformer appears to improve the measured characteristics more neatly and for a wider area in the listening room compared to the classroom.

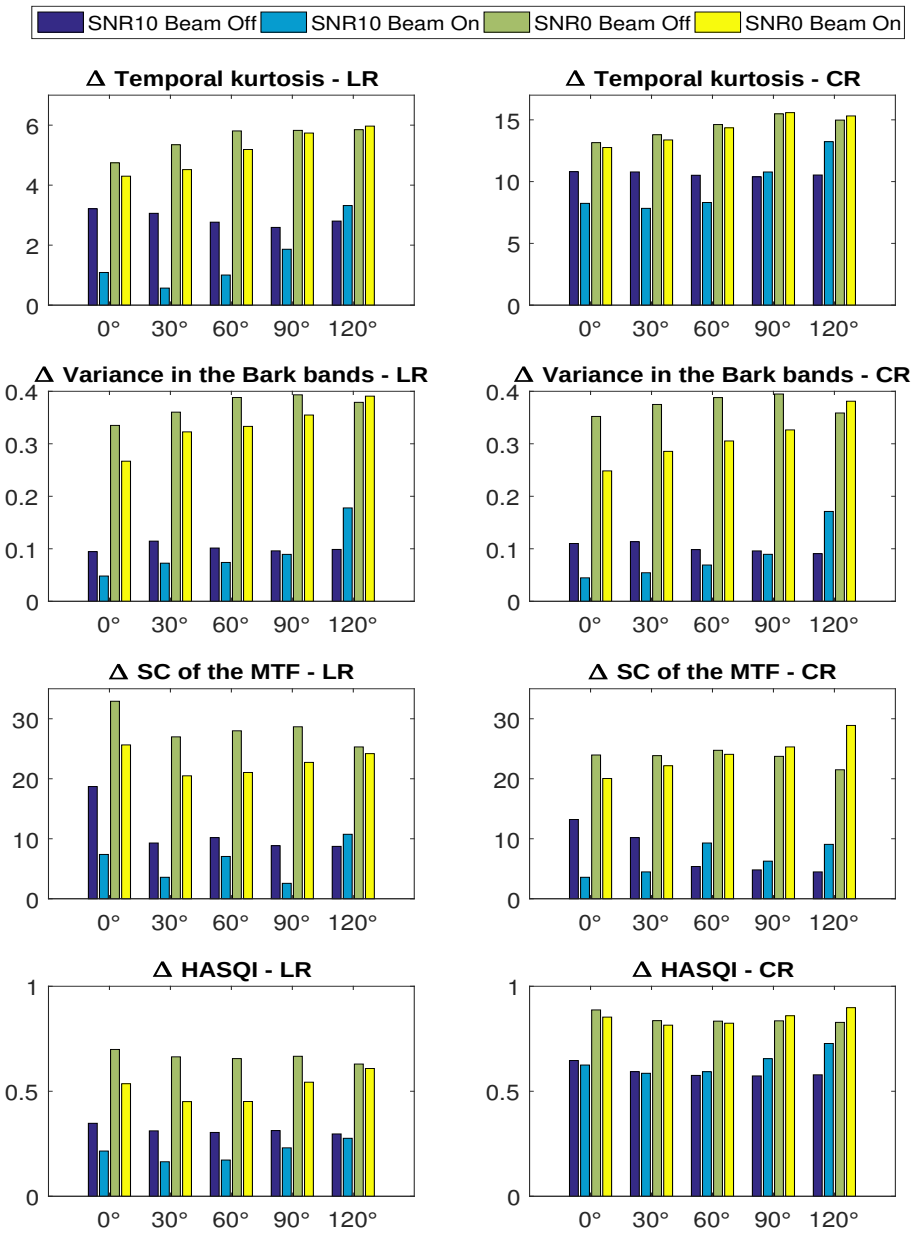

Figure 8: Scores of the objective metrics displayed as a distance to the reference clean signal: temporal kurtosis, variance of energy in the Bark bands (arbitrary scale), spectral centroid (SC) of the modulation function transfer (MTF) and HASQI for SNR = 10 dB and SNR = 0 dB and two environments: Listening room (LR) and Classroom (CR). The lower the bars, the closer to the reference, and consequently the higher the quality.

6. Conclusion

In this study, the effect of a static beamformer on the quality of recorded speech in diffuse babble noise was evaluated. For this purpose, three objective measures, associated with the TFS, spectral content and envelope respectively, were proposed. The reported evaluations show that the use of a static beamformer shifts the studied properties of noisy speech signals closer to those of their clean reference for speech sources located in an area centered in the steering direction of the beamformer. The HASQI indicates that the use of beamforming does not induce speech distortion and might even improve speech quality.

The combination of the proposed metrics could be used in addition to subjective evaluations for the tuning of algorithms aiming to reduce noise and enhance speech intelligibility in hearing device applications. Future works could evaluate the validity of the proposed metrics with a larger number of SNRs and noise types, as well as for different processing of speech enhancement.

Acknowledgement

This study was funded by the Commission for Technology and Innovation (CTI) of the Swiss Federal Department of Economic Affairs, Education and Research (EAER), with grant number 25760.1 PFLS-LS, in collaboration with Sonova AG.

References

[1] B.D. Van Veen, K.M. Buckley: Beamforming: A versatile approach to spatial filtering. IEEE Signal Proc. Mag., Volume: 5, pp 4-24, April 1988.

[2] H. Cox, R. Zeskind, M. Owen: Robust adaptive beamforming, IEEE-ACM T. Audio Spe., Volume: 35, Issue: 10, Oct. 1987.

[3] G. W. Elko, A.-T.N. Pong: A simple adaptive first-order differential microphone. IEEE Work. Appl. Sig., pp 169-172, New Paltz, NY, USA, Oct. 1995.

[4] P. Rakesh ; S. Siva Priyanka, T. Kishore Kumar: Performance evaluation of beamforming techniques for speech enhancement. Fourth Int. Conf. on Signal Processing, Communication and Networking (ICSCN), 2017.

[5] J.G. Beerends, E. Larsen, N. Iyer, J.M. van Vugt: Measurement of speech intelligibility based on the PESQ approach. Proc. Int. Conf. Meas. Speech Audio Quality Netw. (MESAQIN), Prague, Czech Republic, 2004.

[6] A. Spriet, L. Van Deun; K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, J. Wouters: Speech Understanding in Background Noise with the Two-Microphone Adaptive Beamformer $BEAM^{TM}$ in the $Nucleus\ Freedom^{TM}$ Cochlear Implant System. Ear Hearing, Volume: 28, Issue: 1, pp 62-72, Feb. 2007.

[7] G.H. Saunders, J.M. Kates: Speech intelligibility enhancement using hearing-aid array processing, J. Acoust. Soc. Am., Volume: 102, pp 1827, 1997.

[8] M. Kompis, N. Dillier: Noise reduction for hearing aids: Combining directional microphones with an adaptive beamformer. J. Acoust. Soc. Am., Volume: 96, pp 1910-1913, 1994.

[9] T.H. Falk, C. Zheng, and W-Y. Chan: A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE-ACM T Audio Spe., Volume: 18, Issue: 7, Sept. 2010.

[10] T. Neher: Relating hearing loss and executive functions to hearing aid users' preference for, and speech recognition with, different combinations of binaural noise reduction and microphone directionality. Front. Neurosci. 8:391, Dec. 2014.

[11] M.E. Lockwood, D.L. Jones, R.C. Bilger, C.R. Lansig, W.D. O'Brien Jr, B.C. Wheller, A.S. Feng: Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. J. Acoust. Soc. Am., Volume: 115, pp 379-391, 2004.

[12] Y. Hu, P.C. Loizou: Evaluation of Objective Quality Measures for Speech Enhancement. IEEE-ACM T Audio Spe., Volume: 16, Issue: 1, Jan. 2008.

[13] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. IEEE 2011, Int. Conf. Acoust. Spee., Proceedings, Volume: 2, pp 749-752, Nov. 2001.

[14] S. Quackenbush, T. Barnwell, M. Clements: Objective Measures of Speech Quality, NJ, Englewood Cliffs:Prentice-Hall, 1988.

[15] J. Ma, Y. Hu, P. Loizou: Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", J. Acoust. Soc. Am., Volume: 125, Issue: 5, pp 3387-3405, 2009.

[16] J.G. Beerends, C. Schmidmer, J. Berger, M. Obermann,R. Ullmann, J. Pomy, M. Keyhl: Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I-Temporal Alignment. J. Audio. Eng. Soc., Volume: 61, Issue: 6, pp 366-384, June 2013.

[17] T.H. Falk, V. Parsa, J.F. Santos, K. Arehart, O. Hazrati, R. Huber, J.M. Kates, and S. Scollie: Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices. IEEE Signal Proc. Mag., Volume: 32, Issue: 2, pp 114-124, Mar. 2015.

[18] J.M. Kates, K.H. Arehart: The Hearing-Aid Speech Quality Index (HASQI) Version 2. J. Acoust. Soc. Am., Volume: 62, Issue: 3, Mar. 2014.

[19] R. Huber, V. Parsa, S. Scollie: Predicting the perceived sound quality of frequency-compressed speech. PLoS ONE 9(11): e110260. Nov. 2014.

[20] H. Dillon: Hearing Aids. Sydney, NSW: Boomerang Press. pp 218-223 2012.

[21] E. Zwicker: Subdivision of the audible frequency range into critical bands. J. Acoust. Soc. Am., Volume: 33, Issue: 2, pp 248-248, 1961.

[22] T. M. Elliott, F. E. Theunissen: The modulation transfer function for speech intelligibility. PLoS. Comput. Biol. 5(3), 2009.

[23] T. Arai, M. Pavel, H. Hermansky, and C. Avendano: Intelligibility of speech with filtered time trajectories of spectral envelopes. Proc. Int. Conf. Speech Lang. Process., pp 2490-2493, Oct. 1996.