BACKGROUND ADAPTATION FOR IMPROVED LISTENING EXPERIENCE IN BROADCASTING

Yan Tang, Trevor J. Cox, Bruno M. Fazenda *

Acoustics Research Centre
University of Salford, UK

Qingju Liu, Wenwu Wang

Centre for Vision, Speech and Signal Processing
University of Surrey, UK

ABSTRACT

The intelligibility of speech in noise can be improved by modifying the speech. But with object-based audio, there is the possibility of altering the background sound while leaving the speech unaltered. This may prove a less intrusive approach, affording good speech intelligibility without overly compromising the perceived sound quality. In this study, the technique of spectral weighting was applied to the background. The frequency-dependent weightings for adaptation were learnt by maximising a weighted combination of two perceptual objective metrics for speech intelligibility and audio quality. The balance between the two objective metrics was determined by the perceptual relationship between intelligibility and quality. A neural network was trained to provide a fast solution for real-time processing. Tested in a variety of background sounds and speech-to-background ratios (SBRs), the proposed method led to a large intelligibility gain over the unprocessed baseline. Compared to an approach using constant weightings, the proposed method was able to dynamically preserve the overall audio quality better with respect to SBR changes.

Index Terms— Background adaptation, intelligibility, audio quality, listening experience, neural network

1. INTRODUCTION

Speech intelligibility is one of the main issues affecting listeners' experience in TV and radio broadcasts [1]. Low intelligibility is usually caused by factors such as background sound effects, intrinsically unintelligible speech, unfamiliar accents and loud ambient noise in the listening environment.

One approach to enhance speech intelligibility – known as the near-end speech modification (e.g. [2, 3, 4]) – is to alter the speech signal spectrally, temporally or both. For this class of algorithms, it is assumed that the clean speech signal is available while the background (masking) signal is known or can be estimated, but cannot be easily processed. Without changing the speech-to-noise ratio (SNR), some of the modifications were able to boost intelligibility in noise by an amount equivalent to increasing the gain of unmodified speech by more than 5 dB [5]. However, these studies are usually less concerned with the perceived quality of the modified speech. In our recent study [6], it was found that a trade-off between the intelligibility and quality of the modified speech is inevitable, especially in very low SNR conditions. It was also suggested that listening to the modified speech in less adverse conditions can escalate its annoyance to listeners [6].

During the production stage in broadcasting where both speech and background sounds are available, adapting the background sound rather than the speech may be less intrusive to listening experience. In [7], we proposed a method to automatically adjust the background level using an intelligibility model as the perceptual guide for down-mixing in broadcasts. When the background introduces only energetic masking to the foreground speech, this method is able to choose suitable speech-to-background ratios (SBRs) that maintain the intelligibility while providing a good listening experience.

If the level of the background needs to hold constant in order to cater for certain design or artistic purposes, one solution is to apply appropriate near-end modifications to the background signal. Spectral re-weighting [8, 2] is a simple but efficient method to enhance speech intelligibility, provided that the optimal weightings are known. It re-allocates the energy from elsewhere to the frequencies that can be released from masking with additional energy injection. Since the effect of spectral re-weighting, to some extent, is similar to post-filtering (e.g. [9]), it is computationally cheap. In [2], the SNR- and masker-dependent spectral weightings were learnt by maximising an intelligibility model [10] in a closed-loop optimisation procedure, which is however time-consuming, preventing its online operation.

In this paper, we aim to learn the optimal weightings for the background sounds by jointly optimising an intelligibility and an audio quality model. The optimal weightings are then used to train an artificial neural network (NN), which provides a solution for online processing. The performance of the method is finally evaluated by comparing it to the unmodified signals, and the signals modified by static weightings.

*The authors would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

2. METHOD

2.1. Deriving the optimal spectral weightings

After [2], the spectral weightings were optimised for each frequency band on the background sound instead of the target speech signal. The background signal $s(t)$ was disassembled into 34 subbands using gammatone filterbanks, whose centre frequencies $\breve{f}$ span from 100 to 7500 Hz on the equivalent rectangle band scale. The f-th gammatone filter associated with $\breve{f}$ was implemented using a pole-mapping technique [11]:

$$q_f(t) = e^{-j2\pi\breve{f}t}s(t). \quad (1)$$

The output of Eqn. 1 at the f-th frequency band, q_f, was then filtered by a 4th-order filter, whose transfer function H in the Z-domain is,

$$H(z) = \frac{1 + 4az^{-1} + a^2z^{-2}}{1 - 4az^{-1} + 6a^2z^{-2} - 4a^3z^{-3} + a^4z^{-4}}, \quad (2)$$

where

$$a_f = e^{-\mathcal{B}_f \cdot 2\pi/\lambda}, \quad (3)$$

and λ is the sampling frequency. $\mathcal{B}_f$ is the bandwidth of the filter for the f-th band, calculated as,

$$\mathcal{B}_f = 24.7(0.00437\breve{f} + 1) \cdot 1.019. \quad (4)$$

The waveform at the f-th band, s_f, can then be extracted from the filter output u_f,

$$s_f(t) = \frac{(\mathcal{B}_f \cdot 2\pi/\lambda)^4}{3} \cdot \Re(e^{-j2\pi\breve{f}t} \cdot u_f(t)). \quad (5)$$

After applying spectral weighting W_f (in the decibel) to each frequency band, the modified background signal s' was reconstructed by summing across the waveforms of all 34 bands, followed by energy re-normalisation to meet the constant input-output energy constraint,

$$s'(t) = k \cdot \sum_{f=1}^{F=34} s_f(t) \cdot 10^{W_f/20}, \quad (6)$$

where k is a scalar to normalise the signal to ensure the same intensity as s.

To learn the optimal W_f for each speech-background pair at a specified SNR[1], an implementation of pattern search algorithm from the MATLAB Global Optimisation Toolbox was used to explore the space of W with 34 elements. All the settings for the optimisation algorithms were the same as those reported in [2]. Taking both speech intelligibility and overall audio quality into account as two perceptual aspects affecting listening experience, a linear combination of the High-Energy Glimpse Proportion (HEGP [10], as in [2]), and the Perceptual Evaluation of Audio Quality (PEAQ [12]), OM, was used as the objective function during the optimisation process,

$$OM = k_{si} \cdot \text{HEGP} + k_{aq} \cdot \text{PEAQ}, \text{ w.r.t } k_{si} + k_{aq} = 1 \quad (7)$$

where k_{si} and k_{aq} are the weights that balance speech intelligibility and overall audio quality, respectively.

Having observed that the dominance of speech quality in listening experience increases over the improvement of intelligibility [6], a two-parameter sigmoid function was estimated for modelling the relationship between perceptual speech intelligibility score (HEGP) and audio quality weights (k_{aq}), as displayed in Fig. 1. The value of k_{aq} as the function of HEGP was fitted from three points which were chosen based on the characteristics of the HEGP measure and findings from previous studies. When HEGP ≤ 0.1, speech is entirely unintelligible [10], hence maximising HEGP is prioritised ($k_{si} = 1 - k_{aq} = 1$) to boost intelligibility to improve listening experience. When HEGP ≈ 0.6, speech is just about fully intelligible, but due to additional listening effort being required, the overall quality may not yet be the only factor affecting listening experience (here k_{aq} was empirically set to 0.7) [7]. When HEGP ≈ 0.7 with even more favourable SBR, audio quality gradually becomes dominant (k_{aq} was set to 0.9) [6, 7].

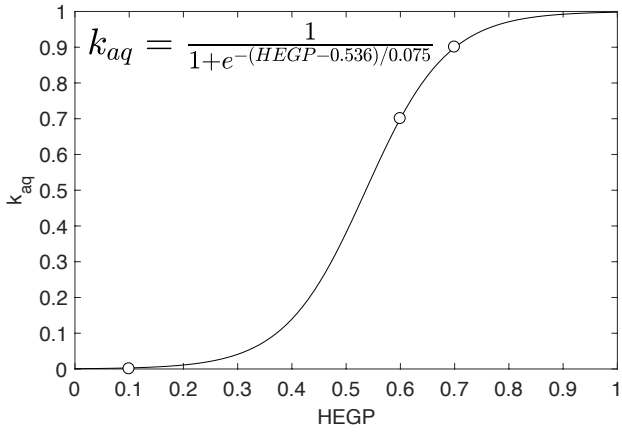

Fig. 1. *Weight of the overall audio quality k_{aq} as a function of HEGP.*

To allow for the fact that the listener in practice hears the speech-plus-background mixture, PEAQ was calculated by comparing the mixture (speech+s' to speech+s) rather than just the background that is being modified (s' to s). This consideration also accounts for the possibility that some artefacts on the background signal due to the modification might be masked by the speech signal in the mix, hence they are not perceptually noticeable to the listener.

[1] Different from [2], in which the weights were sought for the best average performance of a set of stimuli, they were optimised for each speech/masker pair to create the data for NN training.

2.2. Neural network implementation

Closed-loop optimisation is computationally expensive. Performing it for every speech-background pair is impractical for real time applications. A standard feedforward NN was subsequently trained on a limited amount of data, aiming to speed up the processing. The NN consisted of two hidden layers, each of which had the number of neurons that matched the number of the input features. Further tests suggested that increasing the number of hidden layers and neurons did not improve the model performance in this case. For each sample, the mean log-compressed spectra were calculated from every 10ms in all the 34 subbands for both the speech and background signals as the input features. Further including the 34 mean band SNRs, the input feature vector had a total of 102 elements. Other feature combinations were also tested (e.g. only the spectrum vector), but the chosen features led to the best model performance. The outputs of the NN were the 34 optimised weightings. The data was batch-normalised. The tan-sigmoid and linear activation function were used for the hidden layers and the output layer, respectively. The trained NN was finally used to estimate the optimal weightings for adapting the background signal from a given speech-background pair.

3. EVALUATION

3.1. Stimuli

A total of 120 Harvard sentences [13] uttered by a British male talker with a sampling frequency of 16 kHz were used to generate the data for the NN training. Each sentence was mixed separately with a background sound of the same duration that was randomly cropped from cafe noise (CAFE), female competing speech from news broadcasts (CS), crowd noise from a football stadium (CROWDS), a pop song (SONG), the same song but with the vocal being removed (SONG-VR) and classical music (CLASSICAL). Duration of the raw background sound files varies from 4 to 6 minutes. The background signals were downsampled from 44.1 kHz to 16 kHz in order to match the speech signals. Fig. 2 shows the long-term average spectra (LTAS) of the six background sounds.

The SBRs for mixing were from -21 to 9 dB with a 3-dB

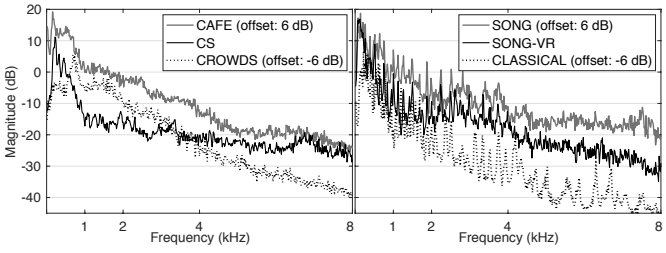

Fig. 2. *LTAS of the background sounds. For illustration purposes, some of the spectra are shifted by an offset of 6 dB.*

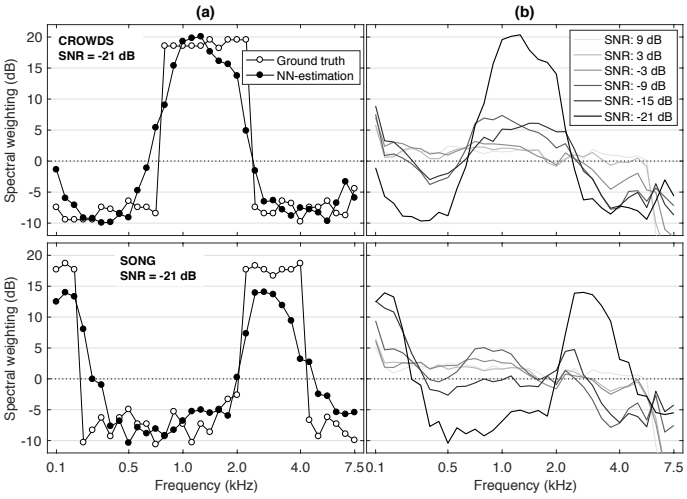

Fig. 3. *(a): Comparisons between optimised (the ground truth) and NN-estimated spectral weightings. (b): NN-estimated weightings at different SBRs.*

increment, leading to a total of 7920 speech-background sample pairs. The optimisation was performed for all samples; the optimised outputs were used as the groundtruth labels for training the NN. During the training, a ratio of 7:3 was used in allocating data for training and validation. The training data was randomised after each epoch.

The plots in column (a) of Fig. 3 compare the optimised and NN-estimated spectral weightings for the same sample in different backgrounds. Despite some minor difference, the overall patterns of the NN-estimated weightings broadly match the ground-truth. Further investigation on the performance measured by HEGP and PEAQ in Table 1 confirms that the NN-estimated weightings achieve similar performance as the ground-truth across SBRs. This result echoes the findings in [2] that the match in general boosting pattern is more important than the consistency in boosting details. Column (b) of Fig. 3 further reveals that the boosting pattern not only varies across maskers, but also across SBRs. With an increase in SBR (from -21 to -15 dB) when the audio quality starts to expedite gaining premium as shown in Fig. 1, the boosting and attenuation become much less drastic, potentially better at retaining the original audio quality.

For evaluation, the speech-background pairs were generated from a set of 300 sentences that did not appear in the NN training. The SBRs ranged from -19.5 to 10.5 dB in the same 3-dB step, in order to inspect the performance of the NN when dealing with unknown conditions. The performance of NN-estimated spectral weightings (as Dynamically-weighted in Fig .4) was evaluated by measuring HEGP for speech intelligibility and PEAQ for audio quality from the adapted background signals. The results from the unmodified signal (as Unmodified) were also presented as the baseline. In addition, the static weightings (as Statically-weighted) proposed in [2] were conversely applied to the background signal (i.e. attenuate the frequencies after 1 kHz for 20 dB) for comparison.

Table 1. Mean difference, $|X_G - X_{NN}|$ and its standard deviation (in the parentheses) in the performance of the ground-truth (X_G) and NN-estimated weightings (X_{NN}), where X is speech intelligibility (HEGP) or overall audio quality (PEAQ) index.

	CAFE	**CS**	**CROWDS**	**SONG**	**SONG-VR**	**CLASSICAL**
HEGP $\in [0,\ 1.0]$	0.03 ($\pm$0.03)	0.02 ($\pm$0.03)	0.04 ($\pm$0.03)	0.03 ($\pm$0.03)	0.01 ($\pm$0.01)	0.04 ($\pm$0.04)
PEAQ $\in [-4.0,\ 0]$	0.27 ($\pm$0.28)	0.26 ($\pm$0.11)	0.48 ($\pm$0.29)	0.56 ($\pm$0.32)	0.52 ($\pm$0.32)	0.34 ($\pm$0.22)

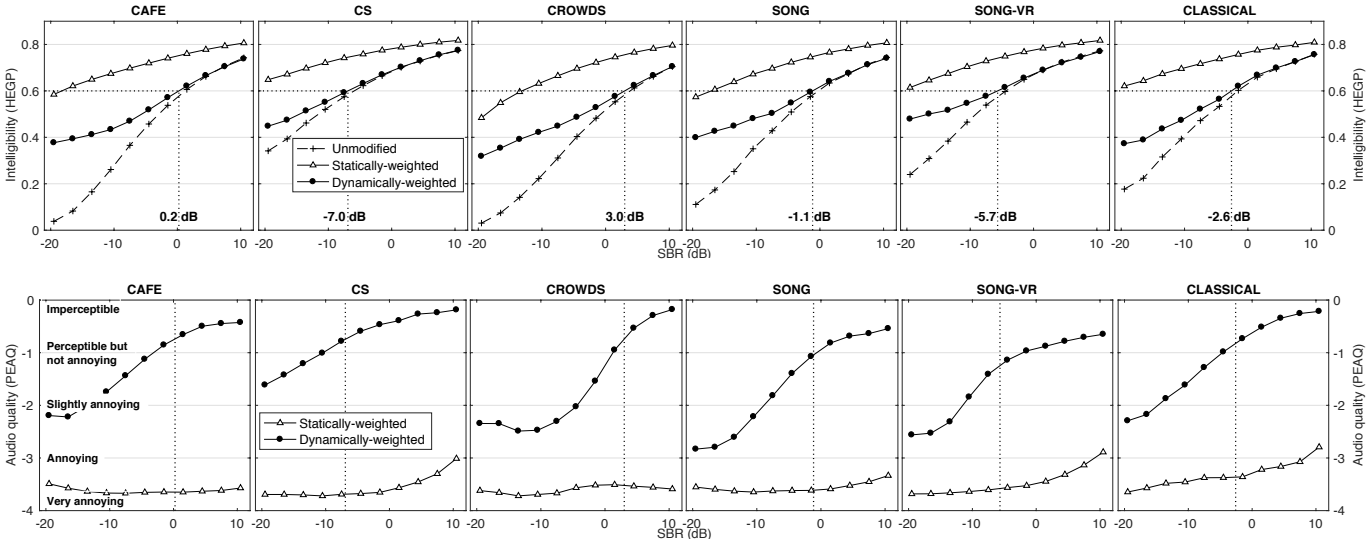

Fig. 4. *Intelligibility (HEGP, upper row) and overall audio quality (PEAQ, lower row) at different SBRs, using unmodified, statically-weighted and dynamically-weighted background sounds. The highlighted SBR in each upper-row subplot led to 0.6 HEGP when the background was dynamically-weighted.*

3.2. Results

Fig. 4 presents the performance of the proposed method in boosting speech intelligibility (upper plots) while maintaining the overall audio quality (lower plots) in the six background sounds. Of all types of backgrounds, "Statically-weighted" always led to the most substantial HEGP gains over "Unmodified" even under extremely negative SBRs (-19.5 dB), at which the speech is entirely unintelligible in some backgrounds (e.g. CAFE and CROWDS). However, this is at a large cost to the overall audio quality – the static weights resulted in PEAQ in the range between "Annoying" and "Very annoying" in almost all conditions.

With "Dynamically-weighted", an improvement in HEGP was achieved. Although the gain is not as large as in the "Statically-weighted" case, "Dynamically-weighted" is able to boost the intelligibility to somewhat intelligible (above 0.3 HEGP) from unintelligible (under 0.1 HEGP) in some of the extreme SBR cases. As the result of the weight allocation mechanism that accounts for the relationship between intelligibility and audio quality, the intelligibility of "Dynamically-weighted" approximately converged to "Unmodified" at the SBRs where 0.6 HEGP is reached. Intriguingly, the PEAQ scores above these SBRs all fall in between -1 (Perceptible but not annoying) and 0 (Imperceptible), except for SONG-VR, indicating excellent overall audio quality. Compared to the PEAQ scores of "Statically-weighted" which stay constantly low over all SBR, "Dynamically-weighted" exhibits a more adaptive manner in catering for both intelligibility and audio quality across different SBRs.

4. CONCLUSIONS

We proposed an approach to enhance speech intelligibility while preserving the overall audio quality for improving listener's listening experience in broadcasting. From jointly optimising an intelligibility and an audio quality measure, the optimised spectral weightings were used to train a NN in order to speed up the signal processing and to deal with unseen situations for practical use. The NN-estimated weightings were then used to alter the energy distribution of the background in the frequency domain. As an adaptive function – which models the relationship between intelligibility and audio quality – was used to determine the weights for the two perceptual factors in listening experience, the optimised spectral weightings are able to keep a reasonable balance between the two factors for the modified background signal. Perceptual listening experiments will be conducted to further validate this method. Another potential extension is to integrate the optimisation stage into the NN training, forming an unsupervised NN training system.

5. REFERENCES

[1] M. Armstrong, A. Brown, M. Crabb, C. J. Hughes, R. Jones, and J. Sandford, "Understanding The Diverse Needs Of Subtitle Users In A Rapidly Evolving Media Landscape," BBC Research & Development White Paper WHP 307, 2015.

[2] Y. Tang and M. Cooke, "Learning static spectral weightings for speech intelligibility enhancement in noise," *Computer Speech and Language*, vol. 49, pp. 1–16, 2018.

[3] T. Zoril, Y. Stylianou, S. Flanagan, and B. C. J. Moore, "Evaluation of near-end speech enhancement under equal-loudness constraint for listeners with normal-hearing and mild-to-moderate hearing loss," *J. Acoust. Soc. Am.*, vol. 141, no. 1, pp. 189–196, 2017.

[4] E. Jokinen, H. Pulakka, and P. Alku, "Phase modification for increasing the intelligibility of telephone speech in near-end noise conditions–evaluation of two methods," *Speech Communication*, vol. 83, pp. 64–80, 2016.

[5] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y Stylianou, V. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[6] Y. Tang, C. Arnold, and T. J. Cox, "A Study on the Relationship between the Intelligibility and Quality of Algorithmically-Modified Speech for Normal Hearing Listeners," *J. Otorhinolaryngol. Hear. Balance Med.*, vol. 1, no. 1, pp. 5.1–5.5, 2018.

[7] Y. Tang, B. M. Fazenda, and T. J. Cox, "Automatic speech-to-background ratio selection for maintaining speech intelligibility in broadcasts using an objective intelligibility metric," *Applied Sciences*, vol. 8, no. 1, pp. 59.1–59.20, 2018.

[8] Y. Tang and Cooke. M., "Optimised spectral weightings for noise-dependent speech intelligibility enhancement," in *INTERSPEECH*, Portland, US, 2012, pp. 955–958.

[9] E. Jokinen, S. Yrttiaho, H. Pulakka, M. Vainio, and P. Alku, "Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech," *J. Acoust. Soc. Am.*, vol. 132, no. 6, pp. 3990–4001, 2012.

[10] Y. Tang and M. Cooke, "Glimpse-Based Metrics for Predicting Speech Intelligibility in Additive Noise Conditions," in *INTERSPEECH*, San Francisco, US, 2016, pp. 2488–2492.

[11] M. Cooke, *Modelling Auditory Processing and Organisation*, Cambridge University Press, 1993.

[12] International Telecommunication Union, "ITU-R BS.1387: Method for objective measurements of perceived audio quality," 2001.

[13] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, pp. 225–246, 1969.