Improvement of broadcast speech intelligibility based on instantaneous listening effort

Rainer Huber, Hannah Baumgartner, Tobias Bruns, Christian Rollwage, Ragini Sinha, Jan Rennies

Fraunhofer IDMT, Division Hearing, Speech and Audio Technology, Oldenburg

Introduction

Depending on the hearing abilities, environment and/or age of TV viewers, the speech intelligibility of dialogs in movies can differ markedly for individual recipients. A possibility to change the intelligibility at home on the TV set to an individual optimal level would be a desirable option for many recipients. An adaptive audio modification would be preferred because a static modification (e.g., a fixed SNR increase) may work in some scenes but be too little or too much improvement in other scenes. This contribution therefore presents a way to provide a second, automatically realized version of a speech/background mix for people with increased demands on intelligibility. The degree of intervention is controlled adaptively and depends on the instantaneously estimated listening effort in relation to a pre-defined target value or an individually desired listening effort. To this end, a single-ended (i.e., reference-free) method is used to evaluate the instantaneous listening effort and to inform a subsequent stage of blind source separation and dialog enhancement. The latter can be configured to keep the listening effort below a specified maximum value and can hence be adjusted to improve the speech intelligibility of broadcast audio for individual needs or different target groups.

Method

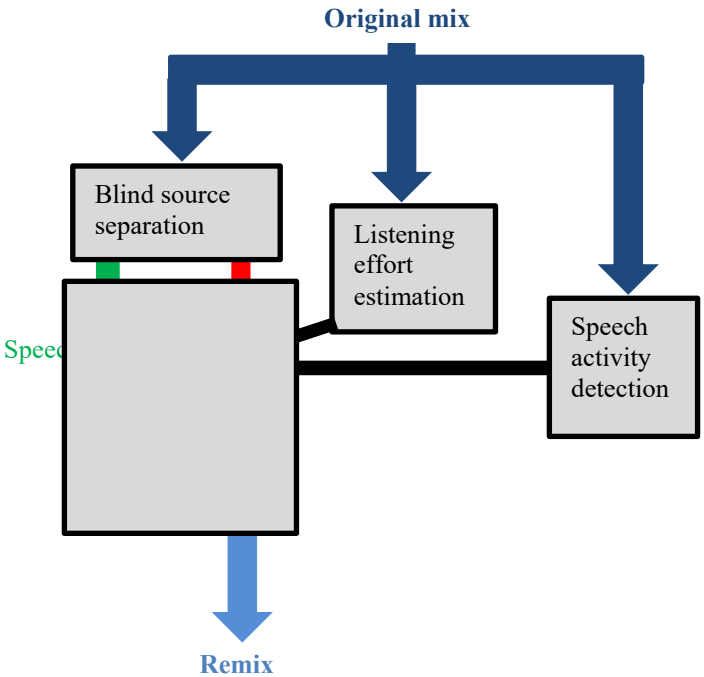

Figure 1: Block diagram of the method

The general approach of the method is depicted in Figure 1. A blind source separation splits the original audio mix into speech and background estimates, which are remixed at a higher speech-to-background ratio in the following. The remixing ratio is determined by the current listening effort of the original mix. The instantaneous listening effort is estimated by a single-ended model. In parallel, a speech activity detector runs on the original mix. In the remixing process, the background signal is attenuated by a factor that is determined by the result of the speech activity detection and the listening effort estimation. This factor is smaller than 1 only if speech activity is detected in the current time frame and if the listening effort exceeds a certain threshold, which can be set according to the individual needs or preferences of the user. The attenuation factor is the smaller, the more the listening effort exceeds the user-defined threshold. The blind source separation, the listening effort estimation, the speech activity detection and the remixing process will be described in more detail in the following.

Blind source separation

The blind source separation (BSS) method was adopted from [1], where it was shown that significant reductions in listening effort can be achieved for broadcast audio material in normal-hearing listeners. It is based on a unidirectional recurrent neural network (RNN) with three long short-term memory (LSTM) layers, each layer having 600 units. The LSTM layers are followed by a fully connected layer with 514 units and a subsequent sigmoid activation function. The networks predicts time-frequency (TF) masks with 257 frequency bins for the speech and the background components of the mix. Each mask is multiplied with the TF representation of the mix to give the TF representations of the speech estimate and the background estimate, which are transformed back into the time domain. LSTMs were chosen to make the network unidirectional and thus causal, and to enable a frame-wise processing, which is needed for real-time applications. A frame size of 32 ms with a 16 ms shift is used.

The network was trained using the Music Speech and Noise (Musan) corpus [2]. This corpus contains 60 h of speech, 42 h of music and 6 h of noise at 16 kHz sampling frequency, single-channel. Despite the fact that TV broadcast audio contains frequencies above 8 kHz and can be stereo, this training corpus was chosen because no publicly available sufficient corpus at a higher sampling rate and in stereo was found. As a work-around for dealing with broad-band, stereo TV audio signals, both channels of the stereo signals are processed by the BSS algorithm separately and a high-frequency extension is applied in the subsequent remixing process (see below).

Listening effort estimation

The method of single-ended listening effort estimation and its application for TV audio signals was presented in [3] and shall thus only be described briefly here. The basic approach is to use a part of an automatic speech recognition (ASR) system, i.e., the processing up to the deep neural network (DNN) which decides for the phoneme classes. Disturbances of speech such as distortions or background noise lead to an increased recognition uncertainty of the ASR system, similar to human speech perception. This uncertainty can be detected and quantified in the recognition system as follows: The DNN produces phoneme posterior probabilities ("posteriorgrams"). A posteriogram represents the temporal course of the probability for the activity of individual phonemes. Disturbances of speech lead to the posteriorgrams being "smeared". The degree of smearing is quantified by a mathematical measure, the “mean temporal distance” or “*M*-Measure”, proposed by Hermansky et al [4]. This measure is used as a predictor of listening effort. In [3], it could be shown that the *M*-Measure correlates very well with mean subjective listening effort ratings of more than 400 TV audio samples (Pearson correlation $r = 0.92$).

For the real-time implementation of the *M*-Measure, a moving average is calculated over a running window of 32 posteriorgram frames (frame length = 30 ms) and a window shift of one posteriorgram frame. The resulting *M*-Measure is then transformed into listening effort scale values by a linear mapping function based on extensive psychoacoustic evaluations. The listening effort scale ranges from 1, meaning “no effort”, to 13, meaning “extreme effort”.

Speech activity detection

The speech activity detection (SAD) algorithm used here was optimized for the application on TV audio signals [5]. It consists of a fully connected feed-forward DNN with 3 hidden layers and 480 neurons per hidden layer. The activation function in each layer is based on a 2-norm, which results in a dimension reduction from 480 to 96 nodes [6]. At the input of the DNN, Mel-Frequency Cepstral Coefficients (MFCCs) are used as feature vectors. The output layer of the DNN consists of two neurons indicating speech or non-speech, respectively, for each input vector. MFCC features cannot capture the temporal structure of speech per se, since MFCC feature vectors consist of relatively short time blocks (approx. 25 ms). For a reliable recognition of speech sounds, the analysis of amplitude modulations is essential [7], for which time windows of approximately 200 ms or more are required [8]. In order to capture a larger temporal context with the DNN-SAD system, the MFCC feature vectors are combined over a time window of 470 ms to form a larger feature vector, before being fed into the DNN. In this way, both temporal and spectral characteristics of the audio signal can be analyzed by the DNN, so that the discrimination between speech and non-speech is possible.

The DNN training was similar to the layer-wise back-propagation algorithm [9], in which each DNN layer is first randomly initialized to build the DNN layer by layer and trained using stochastic gradient descent (SGD) [10]. The training data consisted of approx. 130h of speech and 72h of non-speech. In addition, 170 minutes of TV audio material were used.

Remix stage

In the remix stage, the speech and the background estimate extracted by the BSS are remixed at a higher speech-to-background ratio, if the current time frame contains speech and if the current listening effort estimate exceeds a certain, user-defined threshold. Since the raw listening effort estimates change too quickly, depending on the current phoneme content of the running posteriorgram window, they cannot be directly used to control the gain, because they might be too abrupt and sound unnatural to the listener. To level out these quick changes, again a moving average window with a window length of 3 seconds (evaluated by broadcasting engineers) is applied.

As an alternative to broadband remixes, also frequency-dependent gains can be applied. For this, the speech and background signals are split into 8 bands by an FFT-based filter-bank. The power level of each band is then calculated over each block and exponentially smoothed over time. To calculate the band-dependent gains, the differences between these smoothed band levels of speech and background are then multiplied with a band weighting function adopted from the Speech Intelligiblity Index (SII) [11], corresponding to an overall target speech-to-background ratio.

The high-frequency extension reintroduces the missing high-frequency content caused by the BSS by copying the corresponding band from the original signal mix to the remix of the BSS output. To account for the attenuation of the background component in the remixing process, the original mix is attenuated by a few dB, before copying its high-frequency part of the spectrum. The exact amount of attenuation depends on the attenuation factor for the background signal.

Summary

This contribution presented a flexible framework for enhancing speech intelligibility and listening effort in broadcast applications. It contains three algorithmic core components, i.e., a speech activity detection, an estimation model of perceived listening effort, and blind source separation, which inform and perform a segregation and remixing of the target dialog and the background signals. Each of these components was optimized for applicability in broadcast applications and evaluated in previous studies. Because the entire processing is causal, the framework can be implemented for online processing of audio streams, e.g., to create an alternative broadcast audio track with enhanced speech intelligibility. One major strength of the framework is that it can operate in different modes. In its adaptive mode, the degree of algorithmic intervention is automatically controlled by estimates of speech activity detection and listening effort. The main advantage of this mode is that the original audio mix is only modified at places where listening effort is estimated to be too high and is left unmodified otherwise. This means that the originally intended sound character of the audio stream is maintained as much as possible. To achieve this, a certain amount of processing

time is required to obtain reliable estimates of listening effort and speech activity, which means the algorithmic delay is larger than in the non-adaptive mode. To reduce processing time, the controlling processing blocks of listening effort estimation and speech activity detection can be left out, which sacrifices the adaptivity of the processing, but enables a close to real-time processing with algorithmic delays of a few tens of milliseconds in the current implementation. In this fast version of dialog enhancement, a static background attenuation is specified and the applied to the input stream independent of the actual listening effort. Because of its modular structure, the framework can be easily optimized based on future improvements of individual processing blocks. Possible applications range from broadcasting an additional audio track in TV or radio productions to individualized, adaptive sound processing in streaming application or in end devices at the location of the listener (like TVs, PCs, smartphones, soundbars, or headphones). The latter approach could additionally provide opportunities to combine the proposed dialog enhancement with further sound personalization as, e.g., presented in [12], which could potentially maximize the benefit for the individual user.

References

[1] Westhausen, N. L., Huber, R., Baumgartner, H., Sinha, R., Rennies, J., Meyer, B. T.: Reduction of Subjective Listening Effort for TV Broadcast Signals with Recurrent Neural Networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing, under revision.

[2] Snyder, D., Chen, G., Povey, D.: MUSAN: A Music, Speech, andNoise Corpus, 2015, arXiv:1510.08484v1

[3] Huber, R., Baumgartner, H., Rollwage, C., Goetze, S., Rennies, J.: Erfassung der Höranstrengung fertiger TV-Mischungen, In: Fortschritte der Akustik, DAGA 2019, DEGA, Berlin.

[4] Hermansky, H., Variani, E., Peddinti, V.: Mean temporal distance: Predicting ASR error from temporal properties of speech signal. In Proceedings ICASSP 2013, 38th IEEE Int. Conf. Acoust. Speech Signal Process. doi: 10.1109/ICASSP.2013.6639105.

[5] Moritz, N., Drefs, J., Baumgartner, H., Rennies, J.: Sprachaktivitätserkennung basierend auf Deep Neural Networks für Anwendung in Film und Fernsehen. In Fortschritte der Akustik. DAGA 2016, S.960-963; DEGA, Berlin.

[6] Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks, in International Conference on Acoustics, Speech, and Signal Processing, Florence, pp. 215-219, 2014.

[7] Elliot, T. M., Theunissen, F. E.: The modulation transfer function for speech intelligibility, PLoS Computational Biology, vol. 5, no. 3, e1000302, 2009.

[8] Moritz, N., Anemüller, J., Kollmeier, B.: An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1926-1937, 2015.

[9] Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks, in Advances in neural information processing systems, vol. 19, Vancouver, pp. 153-160, 2007.

[10] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit, in IEEE Automatic Speech Recognition and Understanding Workshop, Hawaii, 2011.

[11] ANSI: ANSI S3.5–1997. American National Standard Methods for the Calculation of the Speech Intelligibility Index. New York: ANSI, 1997.

[12] Rennies, J., Oetting, D., Baumgartner, H., Appell, J.-E.: User-interface concepts for sound personalization in headphones. In Proc. of AES Conference on Headphone Technology, pp. 2-6. Aalborg, Denmark, 2016.