DIALOGUE ENHANCEMENT OF STEREO SOUND

Jürgen T. Geiger, Peter Grosche, Yesenia Lacouture Parodi

juergen.geiger@huawei.com
Huawei European Research Center, Munich, Germany

ABSTRACT

Studies show that many people have difficulties in understanding dialogue in movies when watching TV, especially hard-of-hearing listeners or in adverse listening environments. In order to overcome this problem, we propose an efficient methodology to enhance the speech component of a stereo signal. The method is designed with low computational complexity in mind, and consists of first extracting a center channel from the stereo signal. Novel methods for speech enhancement and voice activity detection are proposed which exploit the stereo information. A speech enhancement filter is estimated based on the relationship between the extracted center channel and all other channels. Subjective and objective evaluations show that this method can successfully enhance intelligibility of the dialogue without affecting the overall sound quality negatively.

Index Terms— Speech enhancement, dialogue enhancement, voice activity detection, stereo enhancement, Wiener filter

1. INTRODUCTION

Recent studies show that many people, especially hearing-impaired listeners, have problems in understanding dialogues in TV sound [1, 2]. Although movie soundtracks are normally carefully mixed in order to achieve a good speech intelligibility, problems can still arise in suboptimal listening conditions. To overcome this problem, approaches were proposed which aim at providing the user a control mechanism which allows for improving speech intelligibility. A straightforward method is proposed in [2] for enhancing the dialogue in discrete 5.1 mixes. Based on the assumption that the relevant dialogue is mixed into the center channel, this approach attenuates all non-center channels. A similar approach is proposed in [3]. For high-quality content delivery channels, such discrete multi-channel signals are typically available. For everyday broadcasting and streaming (e. g. YouTube), however, content is typically only available in the form of a stereo downmix which lacks the discrete center channel. In this case, more sophisticated methods for dialogue enhancement are necessary.

1.1. Related Work

Several methods have been developed in order to boost speech components in a stereo signal. In a first step, such methods typically try to regain a center channel from the stereo downmix. For example, a frequency-domain center extraction technique is proposed in [4]. The extracted center channel can then be amplified (in relation to the left and right channel) to boost the center-panned speech components. In [5], a method for frequency-domain upmixing is described which extracts a panning index to identify the various sources in the signal. Other approaches aim at detecting speech components within the mix. In [6], a speech enhancement approach is proposed which detects speech in movies with a pattern recognition method. More dialogue enhancement methods are summarised in [7]. Theoretically, any kind of conventional monaural speech enhancement method could be applied in this scenario. This includes classical methods such as MMSE speech enhancement [8] as well as novel methods using non-negative matrix factorization [9] or deep neural networks [10].

1.2. Contributions

This work proposes a method for dialogue enhancement of stereo signals. The goal is to boost dialogue components in order to improve speech clarity and intelligibility. The proposed method consists of three steps. First, a center channel is extracted from the stereo downmix which contains all components that are present in both channels of the stereo signal. Typically, this includes the dialogue but also other sounds. To attenuate such other sounds, in a second step, the extracted center channel is further processed by a speech enhancement filter. Finally, in a third step, a voice activity detection is executed with the goal to isolate speech components. The extracted speech components are mixed together with the original signals, to retain all non-speech sounds while boosting the speech components. As main contribution, novel methods are proposed for speech enhancement and voice activity detection which particularly address the application scenario and exploit the availability of stereo signals. Efficient speech enhancement is performed with a Wiener filter which is estimated by regarding the extracted center channel as the target signal and all other channels as noise. For voice activity de-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

tection, a computational simple method based on a measure of spectral flux is presented. Subjective and objective evaluations confirm the potential of the proposed method.

The rest of the paper is organised as follows. In Section 2, the employed method for center channel extraction is described. A novel stereo speech enhancement method is proposed in Section 3, followed by voice activity detection in Section 4. The experimental evaluation is discussed in Section 5, followed by some conclusions in Section 6.

2. CENTER CHANNEL EXTRACTION

As dialogue typically occurs in the center channel of a 5.1 mix, it is reflected in the form of a phantom center in the stereo downmix. In addition, the phantom center might contain other sounds, such as step sounds or other sound effects. Therefore, regaining the center channel from a stereo signal is a first step towards extracting the speech components.

In this work, we use an established method for center extraction which is described in detail in [11] and summarized as follows. This method is based on the assumption that the stereo signal L, R is the result of a downmix of an original three-channel signal L_o,C_o,R_o. The original side signals L_o and R_o are assumed to be orthogonal to each other, and the center signal C_o is assumed to be orthogonal to the side signals. The idea is then to reconstruct the original signals as

$$C_e = \alpha \cdot (L + R), \tag{1}$$

$$L_e = L - C_e, R_e = R - C_e, \tag{2}$$

where α is to be optimised such that the constraint

$$L_e \cdot R_e^* = 0, \tag{3}$$

is fulfilled, which means that the reconstructed signals L_e and R_e should be orthogonal to each other. Under these constraints, a solution for α can be derived as

$$\alpha = \frac{1}{2} \cdot \left(1 - \sqrt{\frac{(L_r - R_r)^2 + (L_i - R_i)^2}{(L_r + R_r)^2 + (L_i + R_i)^2}}\right), \tag{4}$$

where L_r and L_i are the real and imaginary parts of the signal L, respectively. Equation (4) is computed in the frequency domain, meaning that the input signals are represented by their FFT components (for simplicity, the same notation as before is used). The value α is therefore computed in every frequency bin of the FFT representation.

The employed method for center extraction can be interpreted geometrically. Obviously, all sources in the original center channel C_o will end up in the reconstructed signal C_e. The same holds for all sources that are hard-panned to the left or the right. The constraint of orthogonal resulting signals L_e and R_e means that sources that are originally panned between the left and right channel are now panned between center and left or center and right, respectively, in the reconstruction. Further processing can now be performed on the extracted center channel, before an output stereo signal is created by downmixing the three channels.

3. STEREO SPEECH ENHANCEMENT

The result of the center channel extraction are the signals L_e, R_e, and C_e, which are used to estimate a speech enhancement filter. As in classical speech enhancement, the signal model $Y = X + N$ is used, where Y is the observed signal, which is the combination of a target signal X and additive noise N. Generally, it is assumed that X and N are uncorrelated. In order to remove the unwanted noise, either the noise N itself or the signal-to-noise-ratio (SNR) $\frac{X}{N}$ need to be estimated. Most classical methods use monaural processing in order to remove the noise signal N.

A classical approach for speech enhancement is to use a Wiener filter when the SNR is known [8]. In this case, the frequency-dependent filter gain is estimated as

$$G = \frac{\frac{X}{N}}{1 + \frac{X}{N}} = \frac{X}{X + N}, \tag{5}$$

where for the signals X and N, a power representation is used. With the estimated filter gains G, the clean signal can be estimated as

$$\widehat{X} = G \cdot Y. \tag{6}$$

The computation of G according to (5) requires knowledge of the a-priori SNR $\frac{X}{N}$, which can be derived with known noise power N.

In order to circumvent the step of noise power estimation, an efficient method which exploits the availability of a stereo signal is proposed to estimate the Wiener filter for speech enhancement. Based on the assumption that all dialogue components are present in the center channel, C_e is regarded as the target signal X and the noise signal N is composed of $L_e - R_e$. With this interpretation, the speech enhancement filter can efficiently be estimated from the powers of the signals C_e and $L_e - R_e$ as

$$G = \frac{P(C_e)}{P(C_e) + P(L_e - R_e)}, \tag{7}$$

where $P(\cdot)$ denotes the power representation of a signal. This filter is applied on the center channel C_e to remove unwanted surround components that are leaked to the center. Furthermore, it was found that the application of the filter on the channels L_e and R_e extracts direct components that are leaked into these channels. Therefore, the estimated filter G is applied on all three channels resulting from the center extraction process.

To further improve the efficiency, the filter estimation can be performed in spectral bands (e. g., on a Mel scale) instead of a detailed computation in all spectral bins resulting from the FFT. For this purpose, the spectral powers are averaged in frequency bands.

The proposed speech enhancement method removes signal components from the extracted center C_e that origin from the original non-center channels L_o and R_o, while non-speech components from the original center channel C_o are not affected by the estimated filter. The main effect of the filter is

to remove non-speech components (such as music) that occur simultaneous to speech. In order to remove non-speech sounds that are mixed into the original center C_o, a method for voice activity detection is applied.

4. VOICE ACTIVITY DETECTION

A simple, efficient method for voice activity detection is proposed in order to retain only speech components in the signal. The method is based on the spectral flux, which measures the temporal variation of the power spectrum. For a frequency-domain signal $X(m,k)$, with m being the time frame index and k being the frequency bin index, the spectral flux is defined as

$$F_X(m) = \sum_k \Big(|X(m,k)| - |X(m-1,k)| \Big)^2, \quad (8)$$

which measures the temporal fluctuations of the spectral magnitude between subsequent time frames. Spectral flux is a well-known indicator for voice activity [12]. Higher values of spectral flux (due to alternations between consonants and vowels) are expected for speech compared to music and other sounds.

To avoid a computational complex statistical classifier to derive a voice activity decision from the spectral flux feature, we employ a normalisation process that directly leads to a voice activity score. Again, the availability of a stereo signal is exploited. The preliminary voice activity score V is computed as

$$V(m) = a \cdot \left(\frac{F_C(m)}{F_C(m) + F_{L-R}(m)} - 0.5 \right), \quad (9)$$

where the spectral flux of the center signal F_C is normalised with the total spectral flux, composed of the spectral flux F_C and the spectral flux of the side signal $L - R$. The parameter $a \geq 2$ can be used to scale the score. Afterwards, $V(m)$ is limited to $V(m) \in [0,1]$, and thus, the result can directly be interpreted as a voice activity probability.

Finally, center extraction, speech enhancement and voice activity detection are combined to produce a stereo output signal. The speech enhancement filter G according to (7) is applied to the signals L_e, R_e, and C_e resulting from the center extraction. From the enhanced signals, a voice activity decision V is computed according to (9). The voice activity score V is used together with the enhanced signals to mix the output signals,

$$C'(m,k) = p \cdot C_e(m,k) + q \cdot V(m) \cdot G(m,k) \cdot C_e(m,k) \quad (10)$$

where p and q are parameters that control the ratio between the original signal (first summand) and estimated speech component (second summand). Output signals L' and R' are obtained accordingly. With the parameters p and q, the composition of the output signal based on the original input signal and the estimaed speech are controlled. For example, setting $p = 0$ and $q = 1$ corresponds to using only the extracted speech component, whereas with $p = 1$ and $q = 1$, the speech components from the input signal are boosted, while all other components are still retained. From the signals L', R', and C', a stereo downmix can be created as an output signal.

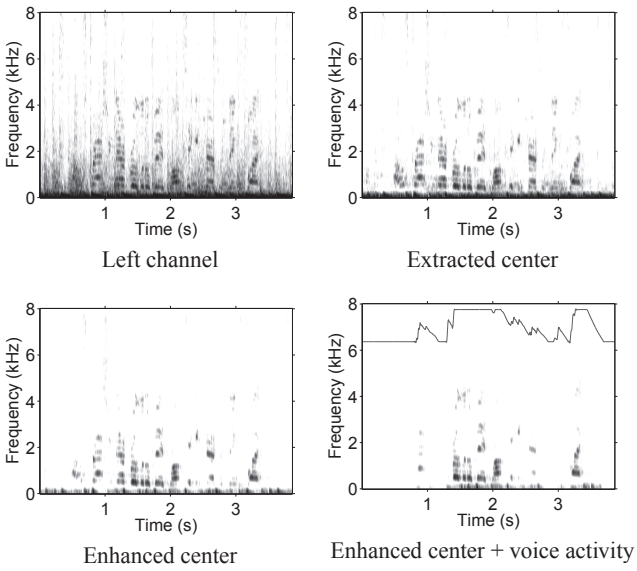

Fig. 1. Spectrogram of different processing steps for a short clip containing speech and background music

Figure 1 illustrates the results of center extraction, speech enhancement, and voice activity detection. For a short extract containing speech and background music, the original left channel, extracted center, enhanced center, and enhanced center combined with voice activity detection are plotted as spectrograms. The last figure also contains the smoothed curve of the voice activity score. These figures show that the proposed method successfully extracts the speech components of the recording.

5. EVALUATION

The goal of the proposed technique is to improve the clarity of the speech component in a stereo mix, under the requirement that no degradation of voice quality should occur. In order to evaluate these aims, subjective and objective evaluations were performed.

5.1. Parametrisation

First, we describe the parameter settings used in the evaluations. Signals are transformed to frequency domain with an FFT, using sine windows with length of 64 ms and 50 % overlap. Several components of the proposed method incorporate temporal smoothing (using the exponential smoothing technique), in order to create smooth output signals and avoid artifacts. In particular smoothing is applied on the numerator

and denominator of (4) and (7) with a smoothing factor of 0.8. The VAD decision (9) is smoothed with an attack smoothing factor of 0.7 and a release factor of 0.98. In order to reduce the computational complexity of center extraction and speech enhancement, the linear frequency scale is transformed with an equivalent rectangular bandwidth filter bank with 43 filters. The parameters p (non-speech gain) and q (speech gain) in (10) are set to $p = 1$ and $q = 1$ to achieve a trade-off between the desired effect of speech boosting and the undesired effect of introducing unpleasant perceptible distortions.

5.2. Subjective Experiments

Clarity of speech (intelligibility) and overall sound quality of the proposed method were evaluated using a 2-alternative-forced-choice procedure. Four different stereo signals containing a mixture of speech, music and background noise were extracted from movies. The signals were then processed with the proposed dialogue enhancement method and compared with the original stereo signal and with an approach using simple center extraction and gain, in which the center is amplified (by 3.8 dB) with respect to the left and right channels.

The stimuli were playback through two typical TV loudspeakers, spanning 20^o and placed 1.6 m from the listener. 13 listeners (1 female, 12 male) between 25 and 40 years participated in the test. The test consisted of 2 independent sessions in which the two attributes were evaluated. All possible pairs were presented twice: once in an AB configuration and second in a BA configuration, giving in total 24 sequences per session. Before each session, a short training was done to help listeners familiarise with the stimuli and the test procedure. The order of the session and sequence presentation was randomized using a Latin-Square design to avoid carry over effects.

The data analysis was done using the Bradley-Terry-Luice (BTL) model [13]. This model makes it possible to extract a ratio scale from pair comparison data. To assess the validity of the ratios, the likelihood of the model is compared with the saturated model that fits the data perfectly using chi-square statistics [14]. The model can be rejected if the p value is less than 10%.

Fig. 2 (left) shows the BTL scores obtained for the clarity test. The goodness of fit of the model $[\chi^2(1) = 0.18, p = 0.6699]$ indicates that the BTL model accounts quite well for the data. In other words, the obtained ratio scale can not be rejected. It can be clearly seen that the proposed method is judged to be significantly clearer than the original stereo and simple center extraction with gain approach. Fig. 2 (right) shows the scale values obtained in the sound quality session. The chi-square statistics $[\chi^2(1) = 0.5361, p = 0.4640]$ indicate that also in this case the model accounts well for the data and the scale values can not be rejected. There is no significant difference in sound quality between the proposed method and the simple center extraction and gain approach. There is however a significant difference between both approaches and

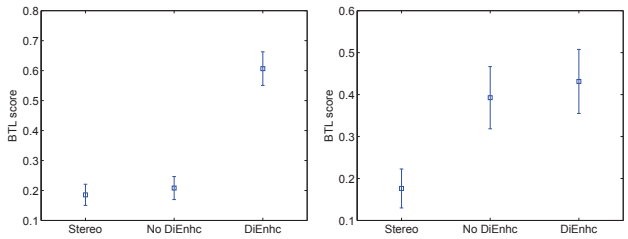

Fig. 2. BTL scores obtained with the speech clarity test (left) and sound quality test (right) and the 95% confidence intervals. Three methods are compared: original stereo, center extraction without dialogue enhancement (No DiEnhc), and center extraction with dialogue enhancement (DiEnhc)

the original stereo. This means that there is a clear preference of the proposed method over stereo, while sound quality is not compromised with the introduction of the dialogue enhancement method.

5.3. Objective Measurements

Two objective measures were used in order to verify the goals of the proposed method. The perceptual evaluation of speech quality (PESQ) measure [15] was used to verify that the proposed method does not introduce any degradations of speech quality. In order to evaluate the potential improvement in speech clarity, the segmental signal to noise ratio (segSNR) measure [16] was used.

The PESQ measure is standardised as ITU-T recommendation P.862. It was designed as an objective voice quality test (with scores between 1 and 5) in telecommunications and measures the distortion of processed speech compared to clean speech. The segSNR measure is a simple time-domain comparison to measure the amount of noise in dB. Higher segSNR values lead to higher listening comfort.

Both evaluation measures require a clean version of the speech signal as a reference. Since the dialogue enhancement method was developed for stereo downmixes of movie soundtracks, the evaluation was carried out with short excerpts from movies. The clean speech component is not available, and therefore the center channel from a 5.1 mix was used as the reference signal. The output of the dialogue enhancement method is a stereo signal, and thus, both of the stereo channels (left and right) are compared to the reference signal for the objective evaluation. The result of both channels is averaged and finally, the average score among all recordings in the test set is computed.

The proposed method is compared to the baseline of a stereo downmix, where no dialogue enhancement or other processing is performed. Furthermore, MMSE speech enhancement according to [8], using minimum statistics noise estimation [17] is used for comparison. In order to produce comparable signals, this speech enhancement method is applied on the left and right channel of a stereo downmix of the

Table 1. Results of the objective measurements

	PESQ	segSNR
stereo	3.00	1.84
MMSE	3.06	3.27
proposed	3.11	3.73
informed downmix	3.70	9.50

test signal, and the estimated clean speech signal is combined with the original left or right channel, respectively. In addition, the measurements were also performed for an informed 5.1 downmix. This downmix follows the recommendation of [2], such that all non-center channels from the original 5.1 signal are scaled by -6 dB prior to the stereo downmix.

As test material, 17 excerpts from Hollywood movies with an average length of 8.5 s are used. All sequences were selected to contain mostly clean speech in the original 5.1 center channel and high amounts of non-speech (music, sound effects) in the other channels.

Objective results are listed in Table 1. Compared to the original stereo signal, both classical MMSE speech enhancement as well as the proposed method achieve a small improvement in terms of PESQ score. This result confirms that the proposed method meets the requirement that no degradation in speech quality should be introduced. Both methods lead to an improvement in segSNR, where the proposed method achieves the best result. The reason for the improved segSNR could be that the proposed method uses the available stereo information in a better way, such that the noise is estimated better for the Wiener filter. The improvement of segSNR obtained with the proposed method, compared to stereo, is almost 2 dB, which shows that the proposed method successfully extracts the speech component from the signal. Compared to the informed downmix, the potential segSNR improvement is by far not fully exploited. However, the informed downmix is favoured in the objective measurements, because the original 5.1 center channel is used as a reference for segSNR computation. The original center contains not always only speech, and some of the contained non-speech components might be removed by the speech enhancement methods, which is punished during the segSNR computation.

6. CONCLUSIONS

We presented a method for enhancing the speech component in a stereo mix. The proposed method consists of extracting a phantom center channel from the stereo signal, followed by novel methods for stereo speech enhancement and voice activity detection. These methods are simple, yet efficient. Subjective and objective evaluations showed that no undesired degradation in speech and overall sound quality are introduced, and confirmed the potential of the proposed method to successfully boost the dialogue component of the signal.

REFERENCES

[1] M. Armstrong, "Audio processing and speech intelligibility: a literature review," *BBC Research & Development Whitepaper*, 2011.

[2] B. G. Shirley, *Improving Television sound for people with hearing impairments*, Ph.D. thesis, University of Salford, 2013.

[3] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue enhancement – technology and experiments," *EBU Technical review*, vol. 2, pp. 1, 2012.

[4] E. Vickers, "Frequency-domain two-to three-channel upmix for center channel derivation and speech enhancement," in *AES Convention 127*, 2009.

[5] C. Avendano and J.-M. Jot, "A frequency-domain approach to multichannel upmix," *Journal of the Audio Engineering Society*, vol. 52, no. 7/8, pp. 740–749, 2004.

[6] C. Uhle, O. Hellmuth, and J. Weigel, "Speech enhancement of movie sound," in *AES Convention*, 2008.

[7] F. Rumsey, "Hearing enhancement," *Journal of the Audio Engineering Society*, vol. 57, no. 5, pp. 353–359, 2009.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014.

[11] C. Brown, "Speech enhancement," 2011, EP Patent 2,191,467.

[12] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. ICASSP*, 1997, pp. 1331–1334.

[13] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324 –345, 1952.

[14] S. Choisel and F. Wickelmaier, "Ratio-scaling of listener preference of multichannel reproduced sound," in *Proc. DAGA*, 2005.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.

[16] J. H. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms.," in *ICSLP*, 1998, pp. 2819–2822.

[17] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001.