NORMALIZATION OF PARTLY OVERLAPPING AUDIO RECORDINGS FROM THE SAME EVENT BASED ON RELATIVE SIGNAL POWERS

Nikolaos Stefanakis[1,2] *and Athanasios Mouchtaris* [1,3]

[1]Foundation for Research and Technology - Hellas, Institute of Computer Science, 70013 Heraklion, Crete, Greece
[2]Technological Educational Institute of Crete, Department of Music Technology and Acoustics Engineering, 74100 Rethymno, Greece
[3]University of Crete, Department of Computer Science, 70013 Heraklion, Crete, Greece

ABSTRACT

Exploiting correlations in the audio, several works in the past have demonstrated the ability to automatically match and synchronize user-generated video or audio files of the same event. Such tools solve for the unknown starting and ending time of each available recording along the event time-line and open the way for collaborative content production approaches. However, a source of difficulty for collaborative processing approaches related to audio is the fact that the different audio recordings may be available at significantly different signal levels. In this paper, we present a normalization approach to automatically define gains for all the recordings so that the variations in the signal levels among different recordings are suppressed. We show that normalization is trivial when all recordings share the same time support but the same process is non-trivial when the recordings partly overlap along time, especially if the acoustic event is characterized by high dynamic variations. We demonstrate the efficiency of the proposed approach under various conditions based on real examples of user-generated audio recordings.

Index Terms— user generated content, audio synchronization, audio normalization, collaborative audio processing

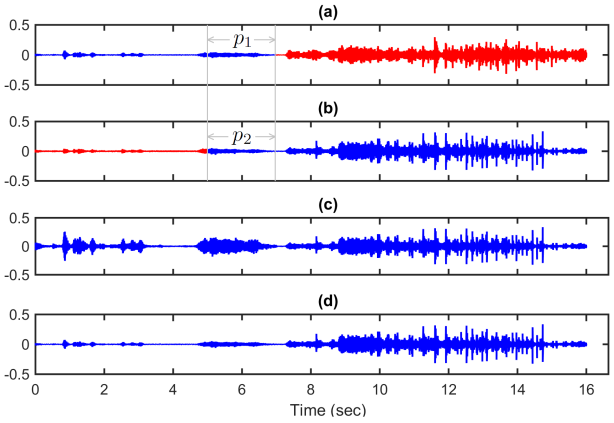

Fig. 1. Two synchronized recordings from the same event are shown in (a) and (b). The result of concatenating the two recordings based on average power normalization is shown in (c) and based on relative power normalization is shown in (d).

1. INTRODUCTION

Given a collection of User Generated audio or video Recordings (UGRs), several approaches have been proposed about how to exploit the available visual and audio content in order to identify video clips associated to the same moment of a public event, to estimate the overlap between these clips and to synchronize them along the same temporal axis. The audio content is a key to solving this problem and several works have shown that the temporal relations between different UGRs can be revealed by exploiting the correlations in their associated audio streams [1–7].

An emerging research challenge is to investigate different means by which this low-quality but organized content can be synergistically processed and combined, so as to produce an improved and more complete audiovisual representation of the captured event. The potential is particularly interesting with respect to the audio modality, as a multitude of synchronized UGRs may be utilized as a multichannel recording of the public event. As shown in [8], simple forms of combination of the different sources of audio content, such as signal superposition and stereo panning, may significantly improve the user experience as opposed to when original UGRs are consumed individually. Additional works demonstrate that the synchronized audio streams can be processed in a more advanced fashion [9–11], so as to enhance the audio components which are common within the different recordings, and to suppress unwanted noise and interference which is unique in each audio capture.

A problem that has not yet been addressed is the fact that in a collection of overlapping UGRs, each recording starts and stops at arbitrary time instants. This means that even if the audio clips are correctly synchronized along the same time-line, the amount of available input audio channels may vary significantly as a function of time. Moreover, different audio recordings are likely to have different signal levels, due to different device specifications and acquisition gains, or due to different distances from the sound sources in the event. All these may complicate the collaborative production process leading, for example, to unwanted transitions in the sound level at the time instant that a certain audio clip starts or stops participating in the mix.

In this paper, we make a first step towards solving this problem by proposing an audio normalization approach with aim to weight the audio recordings, so that their signal levels are consistent with one another and with the dynamics of the actual acoustic event. We demonstrate that this problem is trivial in the case of full overlap, i.e, when all recordings have exactly the same starting and ending times along the time-line, but it is non-trivial in the case of partial

overlap, which is most likely the case in a real situation, since the recordings originate from different users. It is demonstrated that the proposed normalization scheme allows for an automatic adjustment of the audio channel gains even in the case that the acoustic event exhibits significant dynamic variations along time.

2. PROBLEM STATEMENT

We consider a collection of $m = 1, ..., M$ temporally overlapping recordings acquired at the same acoustic event. Using one of the many audio matching and synchronization approaches which are available in the literature, we assume that all M recordings are correctly time aligned along a common time axis. We note that the tools developed in this paper apply only to collections of so-called *connected* recordings, e.g., recordings forming a connected graph. The graph is here implied as follows; each recording represents a node in the graph and two nodes are linked if they temporally overlap for any amount of time. A collection of recordings is connected if the resulting graph is connected, i.e., there is a path connecting any pair of nodes.

Assume also that all recordings are available at PCM format and let $x_m[i]$ denote the value of the ith sample of the mth recording. We refer to normalization as the process of defining a set of M normalization gains $\mathbf{g} = [g_1, ..., g_M]^T$ to scale all recordings according to $\hat{x}_m[i] = g_m x_m[i], \forall m$. As an example to understand how important this requirement is, consider the case of two fully overlapping audio recordings, taken from the same real life public event shown in Fig. 1(a) and (b). It can be easily seen from the corresponding waveforms that the two recordings are correctly aligned along the same time-line and capture the same moment from the public event. Assume know, that only the blue-coloured part of each recording is available, e.g., segment $t \in [0\ 7]$ s with respect to recording 1 shown in (a) and segment $t \in [5\ 16]$ s with respect to recording 2 shown in (b). As each audio clip captures a limited part of the event, it would be advantageous to merge the two recordings in order to create a more complete representation of the captured event. In this example, let this be achieved by combining segments $[0\ 7]$ s from recording 1 with segment $(7\ 16]$ s from recording 2. The problem is that, if the two initial recordings have considerably different signal levels, then the resulting audio stream will be characterized by a sudden level transition at $t = 7$ s. However, jumps in audio level are well known as sources of irritation for the listeners [12]. Moreover, for some collaborative audio production approaches which have been recently proposed [8, 11], it is essential that the instantaneous differences in the signal powers of the different mixture components are small. This motivates the use of some sort of scaling for minimizing the signal level differences between different recordings and below, we present two such candidate approaches.

2.1. Normalization based on average signal power

Under the assumption that the acoustic event is an ergodic process, we can assume that the acquired signals have constant powers along time and we can estimate the average signal level with any size of sample; normalization is in this case trivial and it can be accomplished by obtaining an estimation of the average power of the signal, estimated across the entire duration of each recording. In particular, if N_m is the duration of the mth recording in samples, a normalized version can be obtained through the process $\hat{x}_m[i] = g_m x_m[i]$ with g_m defined from

$$g_m = g_0 / \sqrt{\frac{1}{N_m} \sum_{i=1}^{N_m} x_m^2[i]}, \quad (1)$$

and g_0 related to a reference average signal power. In this paper, we fix g_0 to be equal to the inverse of the square roof of the average power of the first recording. As a consequence, the first element in the resulting gain vector will always be equal to 1. We will refer to this approach as Average Power based Normalization (APN). With respect to the example of Fig. 1, the result of merging the two recordings using APN is shown in subfigure (c). Unfortunately, we can see that the attempt to equalize the average powers of the two audio clips has resulted to an over-amplification of the first audio stream. This is related to the fact that real-life acoustic events are in general not ergodic and may exhibit significant energy variations along time. As a consequence, normalization based on APN is not guaranteed to preserve the dynamic variations of the actual sound scene.

2.2. Normalization based on relative signal powers

Intuitively, a better approach to normalize the two recordings is to consider relative signal powers, by using a measure of the signal energy along the time range where the two recordings overlap. Let p_1 and p_2 denote the signal energy of the first and second recording, respectively, measured along segment $t \in [5\ 7]$ s. If we use recording 2 as a reference ($g_2 = 1$), we can scale the 1st audio stream with $g_1 = \sqrt{p_2/p_1}$ and the result of merging the two recordings is shown in Fig. 1(d). Apparently, this approach better respects the variations in the dynamics of the actual event. However, generalization of this approach to the case of more than two audio recordings is not so trivial, as shown in the section that follows.

3. GENERALIZATION OF RPN

Consider a collection of $M \geq 2$ recordings forming a connected graph. Without loss of generality, we may illustrate the required notations based on the three audio clips shown in Fig. 2. The points in time corresponding to the beginning or ending of each recording define the so-called transition points. We use j to index a time segment extending between two consecutive transition points. Let also $p_{m,j}$ denote the energy of the unscaled mth recording in the jth segment and let $c_j \in \mathbb{N}^+$ denote the plurality of recordings which are active in the jth segment. Finally, we use the notation $\mathbb{S}(m)$ to denote the set with the segment indexes which fall within the range of the mth recording. We may also let $p_{m,j} = 0$ if j is not an element of the set $\mathbb{S}(m)$. Our approach for obtaining normalization gains relies on the assumption that we can find weights $w_1, ..., w_M$ such that

$$\sum_j p_{n,j} w_n = \sum_{j \in \mathbb{S}(n)} \sum_{m=1}^{M} \frac{1}{c_j} p_{m,j} w_m, \ n = 1, ..., M \quad (2)$$

holds. This expresses the belief that the total signal energy in the nth audio stream, scaled by its weight w_n, must equal the sum of the average energy calculated across all segments which belong to the range of the nth recording.

The condition in (2) can be expressed in terms of matrix vector products as

$$\mathbf{U}\mathbf{w} = \mathbf{V}\mathbf{w}, \quad (3)$$

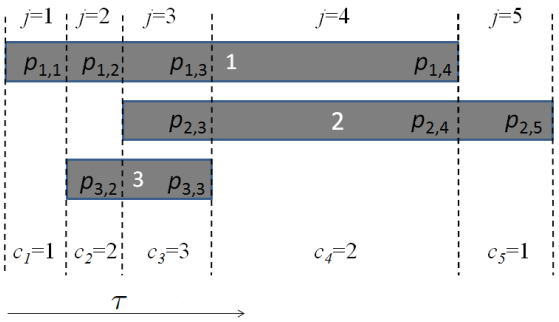

Fig. 2. Three partly overlapping audio recordings defining six transition points and five time segments indexed with j.

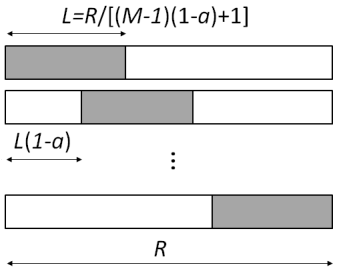

Fig. 3. Using a collection of M fully overlapping audio clips, each clip is deliberately cropped in time, as a function of the overlap parameter a, in order to simulate cases of partly available and partially overlapping clips.

where $\mathbf{w} = [w_1, ..., w_M]^T$ is the vector with the unknown normalization weights, $\mathbf{U} \in \mathbb{R}^{M\times M}$ is a diagonal matrix whose nth diagonal element is equal to $\sum_j p_{n,j}$ and $\mathbf{V} \in \mathbb{R}^{M\times M}$ is a fully populated matrix defined as

$$V_{n,m} = \sum_{j\in\mathbb{S}(n)} \frac{1}{c_j} p_{m,j}. \tag{4}$$

Observe now that the span of all vectors $\mathbf{w}$ satisfying Eq. (3) is actually the nullspace of matrix $\mathbf{Z} = \mathbf{U} - \mathbf{V}$ [13]. Also, it is relatively easy to see that in general $\mathbf{Z}^T\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ and $\mathbf{0}$ are $M \times 1$ vectors full of ones and zeros respectively. This means that the nullity of $\mathbf{Z}$ is at least one[1].

Using any computer program which calculates the null space of a matrix, normalization gains can be thus defined $\forall m$ using $g_m = \sqrt{|w_m^o|}$, where $\mathbf{w}^o = [w_1^o, ..., w_M^o]^T$ is the first (or the one and only) basis vector returned by the program. We refer to this approach as Relative Power based Normalization (RPN) in what follows.

We prove in [14] that in the case of fully overlapping recordings, there is one and only basis $\mathbf{w}$ satisfying $\mathbf{Zw} = \mathbf{0}$ and moreover, in this case the normalization gains obtained based on the APN and the RPN approach, $\mathbf{g}^{AP}$ and $\mathbf{g}^{RP}$ respectively, are collinear. This means that we can always find a scalar $\lambda \in \mathbb{R}^+$ such that $\mathbf{g}^{AP} = \lambda\mathbf{g}^{RP}$ and thus the two approaches are equivalent. However, as the amount of overlap between the recordings decreases, APN and RPN gains may become significantly different. In order to make RPN and APN gains directly comparable to one another, we also scale the RPN gains with an appropriate factor so that g_1^{RP} is also equal to 1.

4. EXPERIMENTAL VALIDATION

This section is devoted for demonstrating the potential of RPN to improve normalization compared to APN in the case of partial temporal overlap. In order to have a basis for evaluating the efficiency of the two approaches, we need a way for defining ground truth normalization gains $\mathbf{g}^{gt}$. Given the fact that when the recordings fully overlap the normalization gains derived by the two techniques are equivalent, we can consider that in the case of full overlap $\mathbf{g}^{gt} = \mathbf{g}^{AP} = \mathbf{g}^{RP}$. Our approach for evaluating the performance is the following; we start with a collection of fully overlapping recordings extending along a common time range R, as shown in Fig. 3, and we calculate the ground truth normalization gains $\mathbf{g}^{gt}$. We then continue by deliberately decreasing the time extend of each recording in order to simulate cases of partially available and partly overlapping recordings. As shown in Fig. 3, different time regions (the gray coloured regions) can be activated for each recording as a function of the parameter $a \in (0, 1)$, which is associated to the amount of overlap. For each value of a, we take into account all the possible permutations of the M recordings, so that the entire time range of each audio clip is used. We then calculate APN and RPN normalization gains as a function of a and we compare the resulting gain vectors directly with $\mathbf{g}^{gt}$ in order to asses Gain Deviation (GD), in dB, defined as

$$GD^{AP} = \frac{1}{M}\sum_{m=1}^{M}\left|20\log_{10}\frac{g_m^{AP}}{g_m^{gt}}\right| \tag{5}$$

for APN and similarly for RPN. Obviously, the closer to 0 that GD is, the better the fit between the gains returned by each approach and the ground truth gains.

As an additional criterion for judging the performance, we examine the degree with which the partially available audio streams and resulting normalization gains track the ground truth energy profile $g^{gt}(\tau)$, defined as a function of time $0 \le \tau \le R$ as

$$e^{gt}(\tau) = \frac{1}{M}\sum_{m=1}^{M} p_m(\tau)(g_m^{gt})^2, \tag{6}$$

with $p_m(\tau)$ a measure of the instantaneous energy of the mth recording at time τ, assuming that each recording is available along its full temporal extent. The restored energy profile is on the other hand defined as the average energy estimated from the set of the partly available and normalized recordings and can be calculated for each value of a as

$$e^{AP}(\tau) = \frac{1}{c(\tau)}\sum_{m\in\mathbb{S}(\tau)} p_m(\tau)(g_m^{AP})^2, \tag{7}$$

for APN and similarly for the RPN technique. In analogy to c_j, now, $c(\tau)$ denotes number of participating clips as a function of τ, $p_m(\tau)$ is the instantaneous energy assuming partially available recordings and $\mathbb{S}(\tau)$ is the set with the indexes of the active audio clips at each time-frame. The Reconstruction Error (RE) defined as

$$RE^{AP} = 10\log_{10}\frac{\sum_\tau \left|e^{gt}(\tau) - e^{AP}(\tau)\right|}{\sum_\tau e^{gt}(\tau)}, \tag{8}$$

represents the error between the ground truth and the restored energy profile in dB. For calculating the instantaneous energy in this paper,

[1]The fact that $\mathbf{1}$ is in the nullspace of $\mathbf{Z}^T$ does not mean that $\mathbf{1}$ is also in the nullspace of $\mathbf{Z}$ since $\mathbf{Z}$ is not symmetric.

we apply a segmentation of the signals using non-overlapping time frames of 0.0427 s length.

In a first series of experiment we use synthetic data based on the audio recording depicted in Fig. 1(a). The particular audio file has 13 s of duration and is characterized by high dynamic variation, representing thus a challenging case for normalization. We produced four replicas of this recording and we artificially contaminated each replica with different amounts of babble noise (crowd noise in particular), ensuring that the noise components are uncorrelated among the different audio streams. This study is interesting as it reveals the theoretical advantage that RPN may achieve against the APN approach, for the case that all recordings share an identical (but possibly scaled) common component. We note however that the defined normalization scheme does not make any distinction between noise and common component; it is assumed that noise contributes to the signal energy in the same degree as the other signal components.

The results are shown in Fig. 4 in terms of GD, in (a), and RE, in (b), averaged across all permutations, as a function of the overlap parameter a ranging from 0.3 to 0.9. It can be seen that RPN may achieve perfect recovery of the ground truth gains in the case of high SNR while its performance degrades as SNR decreases. On the other hand, APN produces large errors as it is very sensitive on the signal energy in each segment. As expected, both APN and RPN performance improves as the overlap factor increases. Also, it is interesting to observe that, RPN and APN have opposite trends with respect to SNR; APN improves with decreasing SNR, which is not surprising considering that the added noise component is more "ergodic" than the common component itself.

A second series of experiments was performed based on real user generated audio recordings which were selected from two different public events. Each collection consisted of $M = 4$ overlapping recordings captured with different devices and with durations ranging from 12.5 to 26 s. Excerpts from different parts of the event were selected so as to enable demonstration of the normalization performance under varying conditions; two of the excerpts are characterized by low dynamic variations (DV), one at high SNR and one at low SNR. The third excerpt on the other hand is characterized by high DV and high SNR. We note here that the characterization regarding SNR is empirical, referring to a subjective assessment on the degree that the common acoustic components within the different recordings are masked by noise, which is unique at each recording device [14].

The results in terms of GD and RE can be seen in Fig. 5. For the case of low DV, both APN and RPN produce satisfactory results, and moreover, APN seems to slightly improve compared to RPN. It can thus be stated that RPN is not expected to provide any significant advantage compared to APN in the case of low DV. Finally, the third acoustic event characterized by large DV presents the most challenging case for both APN and RPN. Similar as in the synthetic experiment, it can be confirmed here that RPN achieves significantly lower GD and ER values compared to APN, especially at small values of the overlapping factor. In total, RPN provides a more reliable approach for normalizing partly overlapping audio signals, performing equally well as APN in case of low DV and considerably better than APN in case of high DV.

The presented approach may theoretically achieve perfect gain recovery in the case that all recordings share a scaled version of the same audio component. However, in many cases portable recording devices incorporate dynamic compression in their processing chain. The fact that dynamic compression is a highly nonlinear signal transformation and the fact that different devices likely incorporate different compression parameters and compression algorithms dictates

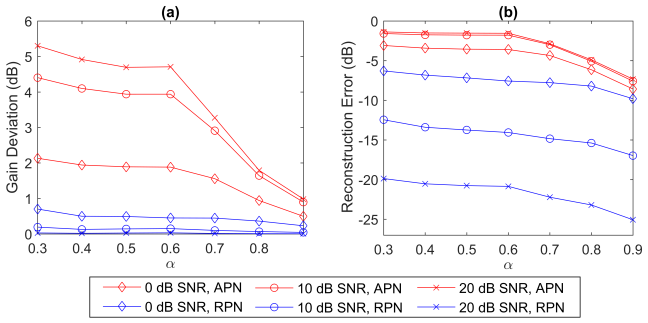

Fig. 4. Gain deviation and reconstruction error as a function of the overlap parameter a for different amounts of babble noise superimposed on the same audio recording.

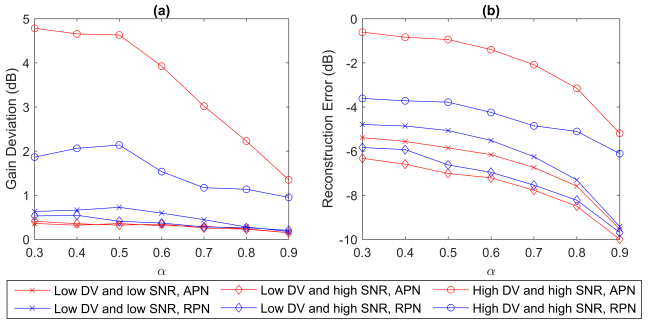

Fig. 5. Gain deviation and reconstruction error as a function of the overlap parameter a for $M = 4$ overlapping recordings taken from three different acoustic events.

that the common components within different recordings are nonlinearly related. Obviously, over-clipping phenomena appearing in some recordings and not in others may have a similar effect. This perhaps also explains why APN appeared to perform slightly better than RPN in two out of three acoustic events used for the experiments in Fig. 5.

As a final conclusion derived from this evaluation, we note that we used MATLAB function *null* for calculating the nullspace of matrix $\mathbf{Z}$. Although we haven not been able to establish a theoretical proof about the uniqueness of $\mathbf{w}^o$ for the case of partial overlap, we observed that the nullity of $\mathbf{Z}$ was always equal to 1. This is probably an indication that uniqueness of the solution for RPN can be proved, but this is in the scope of future work.

5. CONCLUSION

In this paper, we have presented a systematic approach for automatically adjusting the gains of user generated audio recordings, showing that normalization is trivial when all audio recordings share the same time support but becomes a more complex problem when the recordings partly overlap. The proposed approach, RPN, provides a more reliable solution compared to APN, by respecting the dynamic variations that characterize the acoustic event as it unfolds along time. Extension of this work to other metrics of energy as well as to other types of modalities is obvious; for example, in this paper we used signal energy, which is an objective metric, but the approach may be straightforwardly adapted to subjective measures of energy such

as loudness. Finally, it is possible that many other types of signals involving measures of energy in the form of partly overlapping time series may benefit from the proposed normalization approach.

6. ACKNOWLEDGEMENT

The project leading to this application has received funding in part by the European Unions Horizon 2020 research and innovation programme: under grant agreement No 687605, Project COGNITUS, and under the Marie Skłodowska-Curie grant agreement No 644283, Project LISTEN.

7. REFERENCES

[1] P. Shrestha, M. Barbieri, and H. Weda, "Synchronization of multi-camera video recordings based on audio," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 545–548.

[2] L. Kennedy and M. Naaman, "Less talk, more rock: Automated organization of community-contributed collections of concert videos," in *Proceedings of the 18th international conference on World Wide Web*, 2009, pp. 311–320.

[3] C. Cotton and D. Ellis, "Audio fingerprinting to identify multiple videos of an event," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2010, pp. 2386–2389.

[4] S. Bano and A. Cavallaro, "Discovery and organization of multi-camera user-generated videos of the same event," *Journal of Information Sciences*, vol. 302, pp. 108–121, 2015.

[5] J. Bryan, P. Smaragdis, and J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2012, pp. 2389 – 2392.

[6] J. Kammerl, N. Birkbeck, S. Inguva, D. Kelly, A. Crawford, H. Denman, A. Kokaram, and C. Pantofaru, "Temporal synchronization of multiple audio signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2014, pp. 4603–4607.

[7] N. Stefanakis, S. Chonianakis, and A. Mouchtaris, "Automatic matching and synchronization of user generated videos from a large scale sport event," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017.

[8] N. Stefanakis, M. Viskadouros, and A. Mouchtaris, "A subjective evaluation on mixtures of crowdsourced audio recordings," in *Proc. Eur. Signal Proc. Conf. (EUSIPCO)*, 2017.

[9] M. Kim and P. Smaragdis, "Collaborative audio enhancement using probabilistic latent component sharing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2013, pp. 896 – 900.

[10] M. Kim and P. Smaragdis, "Efficient neighborhood-based topic modelling for collaborative audio enhancement on massive crowdsourced recordings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2016.

[11] N. Stefanakis and A. Mouchtaris, "Maximum component elimination in mixing of user generated audio recordings," in *Proc. Multimedia Signal Processing (MMSP)*, 2018.

[12] "Loudness normalisation and permitted maximum level of audio signals," https://tech.ebu.ch/docs/r/r128.pdf.

[13] S. Roman, *Advanced Linear Algebra*, Springer, 2005.

[14] N. Stefanakis and A. Mouchtaris, "Supplementary material," http://users.ics.forth.gr/nstefana/icassp2018.