A Subjective Evaluation on Mixtures of Crowdsourced Audio Recordings

Nikolaos Stefanakis*, Menelaos Viskadouros† and Athanasios Mouchtaris*,†
*FORTH-ICS, Heraklion, Crete, Greece, GR-70013
†University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-70013

***Abstract*—Exploiting correlations in the audio, several works in the past have demonstrated the ability to automatically match and synchronize User Generated Recordings (UGRs) of the same event. Considering a small number of synchronized UGRs, we formulate in this paper simple linear audio mixing approaches to combine the available audio content. We use data from two different public events to perform a comparative listening test with the goal to assess the potential of such mixtures in improving the listening experience of the captured event, as opposed to when each UGR is consumed individually. The results of the listening tests indicate that, even with just a small number of overlapping UGRs, the outcome of the mixing process gains higher preference in comparison to original UGRs played back individually.**

I. Introduction

We live in the era of portable multimedia devices, drones and smartphones, devices capable of capturing every moment of our lives and of the public events that we attend. Audiovisual recordings from these devices, produced by users attending the same public event, become available through the social media and the large number of websites which provide video and audio content. The availability of such massive amount of User Generated Recordings (UGRs) has triggered new research directions related to the search, organization and management of this content, and has provided inspiration for new business models for content storage, retrieval and consumption.

Given a collection of UGRs, several approaches have been proposed about how to exploit the available visual and audio content - as well as several types of metadata - in order to identify video clips associated to the same moment of the captured 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 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 improve both audio and visual aspects of the captured public event (see references in [6] for applications related to visual content). This potential is examined in this paper from the perspective of the audio modality, by utilizing a multiplicity of synchronized UGRs as a multichannel recording of the acoustic event. We refer the reader to the works of Kim et al in [8], [9], as one of the earliest approaches in how a multitude of synchronized UGRs can be exploited for producing a single audio stream with improved properties compared to its constituent parts.

Considering a small collection of overlapping UGRs which are synchronized along a common time axis, in this paper we investigate simple linear mixing approaches for constructing a monophonic or stereophonic audio stream which combines all the available footage at each time instant. We then perform a comparative listening test with the goal to assess the potential of such mixtures in improving the listening experience of the captured event, as opposed to when each UGR is consumed individually. To our knowledge, it is the first time that such a subjective assessment is applied on data acquired from real-life public events. The results of the listening tests indicate that even with just a small number of overlapping UGRs, the outcome of the mixing process gains higher preference in comparison to when the constituent UGRs are played back individually.

II. Mixing techniques

Consider a collection of M UGRs, available at common PCM format and sampling rate F_s, which are synchronized with one another and properly aligned along the same time axis [7]. Also, assume that all these recordings fully overlap along a continuous time interval from time t_{start} to time t_{end}. Imagine now that along this time segment, we would like to exploit the available content in order to provide an enhanced acoustic representation of the event. The first thing that we propose to do is to normalize the recordings with respect to a target average power. Working with normalized versions, rather than with the original versions, ensures that each audio recording has equal significance in the mixing process. This may to some degree prevent, for example, recordings which are acquired at a small distance from the main acoustic sources to mask those which are acquired at distances further apart. In this paper, normalization is accomplished by obtaining an estimation of the average power of the signal, estimated across the entire duration of each UGR. In particular, if we let $x_m[n]$ denote the nth sample of the mth UGR and if N_m is its duration in

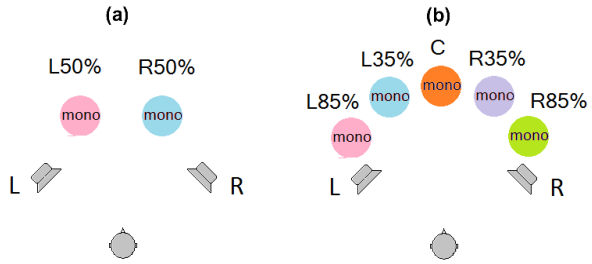

Fig. 1. Synthesizing a *stereo2* version from two *mono* UGRs in (a) and a *stereo5* version from five *mono* UGRs in (b).

samples, a normalized version is obtained through the process $\hat{x}_m[n] = \mu_m^{-1} x_m[n]$ with μ_m defined as

$$\mu_m = \sqrt{\frac{1}{N_m} \sum_{n=1}^{N_m} x_m^2[n]}. \quad (1)$$

Intuitively, the simplest approach to combine the content is to superimpose the recordings. In a monophonic setting, this can be mathematically expressed as

$$s[n] = \sum_{m=1}^{M} \hat{x}_m[n + L_m], \;\; n = 1, ..., N_p \quad (2)$$

where $N_p = floor(\frac{t_{start} - t_{end}}{F_s})$ is the length of the requested action in samples and L_m is the sample shift required in order to align the mth UGR, so that $\hat{x}_m[1+L_m]$ corresponds to time t_{start}. Note that L_m can be separately calculated for each UGR, based on one of the many synchronization approaches that have been proposed in the literature [1]–[7].

The previous approach can be extended to the case that more output channels need to be served, e.g., for stereophonic reproduction. Amplitude panning is one of the simplest mechanisms for synthesizing a stereo signal, by assigning different weights to the different recordings as

$$s_L[n] = \sum_{m=1}^{M} w_{L,m} \hat{x}_m[n + L_m], \quad (3)$$

$$s_R[n] = \sum_{m=1}^{M} w_{R,m} \hat{x}_m[n + L_m] \quad (4)$$

where $w_{L,m}, w_{R,m} \in [0, 1]$ are the so-called panning weights associated to the left and right channel, $s_L[n]$ and $s_R[n]$ respectively. Typically, these weights are adjusted so that $w_{L,m}^2 + w_{R,m}^2 = 1$ holds.

We note that the presented mixing approaches consider that the M UGRs fully span the specified time extent, i.e., the time of initiation and ending for each one of the original UGRs is before t_{start} and after t_{end}. The more difficult problem, where different UGRs may start or stop within the studied time interval, is outside of the scope of this investigation.

III. Data preparation

Two different datasets were used for evaluation, one generated from users attending a open-door musical concert and one from users attending a football match taking place in a crowded open stadium. The exact process for acquiring the recordings in each event as well as more details about the events and the locations of the participants can be found in [10], while the audio datasets themselves are accessible for direct download in [11]. We note that the audio recordings were acquired in a way that ensured the existence of temporally overlapping content, covering several pre-defined parts of each event. All the recordings were captured with smartphone devices, excluding recordings made with a GoPro device in the case of the concert. All the UGRs were available in compressed format and audio was originally acquired at a sampling rate of 44.1 or 48 kHz. For further processing, each audio stream was converted to PCM format at 48 kHz sampling rate and 16 bit of word-length. An audible artifact appearing regularly in the concert recordings was overclipping, while stadium recordings suffered mostly from wind noise. Recordings originating from the same part of each event were manually identified and stored into the same folder. In each folder, 3 mono and 2 stereo UGRs of reasonable quality were selected for evaluation. For each stereo UGR, an additional mono version was constructed by extracting the left channel only. Thus, in total, 5 mono and 2 stereo UGRs were available for each event excerpt that was used for evaluation.

We now briefly describe the process followed for extracting the synchronization times which were required for the time-alignment of the UGRs. Following previous works dealing with the same problem, we extracted audio fingerprints from each UGR. Based on the extracted fingerprints, cross-correlation was then used in order to estimate the time-shift between UGR pairs. For the case of the concert collection, we were able to reliably detect pairwise time-offsets using the fingerprinting technique described in [12]. On the other hand, for the athletic event, we used a recently proposed type of fingerprint proposed in [7]. The time-offsets required for pairwise synchronization were then easily calculated by detecting the peaks in the values of the generalized cross-correlation [7].

In general, a collection of M recordings involves $M(M-1)/2$ pairwise combinations and corresponding time-offsets. These are a lot more than the minimum number of $M-1$ time-offsets which are actually required for synchronizing the entire collection. The approach followed here was to synchronize all the recordings with respect to a single reference audio recording, and in particular, with the UGR which exhibited the stronget correlations with all the others in the same group. To follow such an approach, we summed together the match-strengths between each audio recording and the $M-1$ others, with match-strength defined as the maximum value of the cross-correlation [7]. The recording with the largest sum was then selected as the reference UGR in each folder. The correctness of the synchronization process was easily confirmed by simultaneous playing back of the synchronized recordings. We also note that audio files were synchronized along a linear time grid of 5 ms and 10 ms resolution for the concert and the athletic event respectively, corresponding

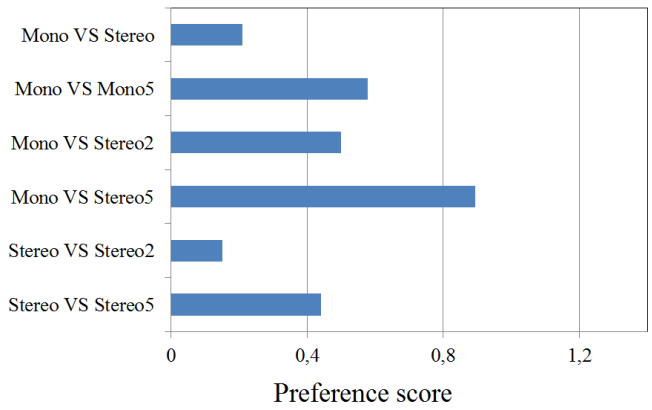

Fig. 2. Preference scores for each comparison case by accounting for all subjective responses in the musical and athletic event.

basically to the hop-size which was used for time-frequency analysis when constructing the fingerprints.

IV. Questionnaire design

The purpose of this evaluation is to answer the following question; In what degree can the previously described mixing approaches improve the impression of an acoustic event transmitted to the listener, as opposed to when each original UGR is consumed individually? An obvious approach for evaluating this potential would be to ask subjects to express their satisfaction with respect to exemplary audio files which are played back one by one. Rather than following this approach, we decided to perform comparative listening tests, where each subject indicates his preference with respect to two different versions of the same acoustic excerpt; one derived directly from the original UGRs, and one produced as a combination of different UGRs, using the mixing techniques described in section II.

In order to cover different formats with which original UGR becomes available as well as different possible mixing approaches, we defined two so-called *original* UGR classes and four *mixed* UGR classes. The original UGR classes were, namely, the *mono* and the *stereo* class, representing the typical audio formats (mono or stereo) with which UGRs become available when directly imported from a portable electronic device. On the other hand, employing different mixing approaches, we decided to construct four so-called mixed classes. Namely, these were

- *mono5*: five normalized monophonic UGRs superimposed using Eq. 2,
- *stereo2*: two normalized monophonic UGRs panned 50% left ($w_{L,m} = 0.924$ and $w_{R,m} = 0.383$) and 50% right, as shown in Fig. 1(a) and
- *stereo5*: five normalized monophonic UGRs panned 85% left ($w_{L,m} = 0.993$ and $w_{R,m} = 0.117$), 35% left ($w_{L,m} = 0.873$ and $w_{R,m} = 0.489$), at the center ($w_{L,m} = 0.707$ and $w_{R,m} = 0.707$), 35% right and 85% right, as shown in Fig. 1(b).

Each question was presented to the subjects by providing two audio streams for comparison, one from the original UGR classes and one from the mixed UGR classes and asking the subject about his/her personal preference. In particular, the cases considered were *mono* vs *mono5*, *mono* vs *stereo2*, *mono* vs *stereo5*, *stereo* vs *stereo2* and *stereo* vs *stereo5*. A final pair considered was *mono* vs *stereo*, providing the only exception of audio versions both originating from the original UGR classes. This was done in order to assess the advantage of a stereophonic recording capability as opposed to monophonic recording which still remains the basic type of recording format for most portable electronic devices.

The two audio files were normalized so as to be perceived equally loud, using the toolbox provided in [13]. Subjects were free to listen to the audio files in each question as many times as they wanted, using headphones, in order to indicate their personal preference given the options “A better than B”, “A slightly better than B”, “ no preference”, “B slightly better than A” and “B better than A”. With respect to this order, the scores assigned to A (or B) were 1 (or -1), 0.5 (or -0.5), 0, -0.5 (or 0,5) or -1 (or -1). The association between audio A and B and each one of the comparison classes was of course random and varied from one question to the other. As both audio streams where produced from exactly the same part of the event, differences in the content between the two recordings were minimized, making it easier for the listener to focus on qualitative aspects of the recordings. This way, we believe that the comparative listening test avoided a great source of bias that would be introduced by the fact that subjects’ expectations with respect to quality and content differ significantly from one listener to the other as well as from one type of public event to the other.

For the athletic event, we selected four short duration excerpts with the aim to cover different types of acoustic content; two excerpts containing chanting of the crowd, one containing clapping from the crowd and one segment without any distinct crowd activity. The average duration of these excerpts was approximately seven seconds. With respect to the concert, four excerpts from four different songs were selected, of approximately six seconds of duration each. All the audio samples were provided for playback in PCM format at 44.1 kHz sampling rate.

In total, 8 different questionnaires were constructed. Four questionnaires with content only from the concert and four with content only from the athletic event. In each case, the four excerpts were distributed with a random order across the four different questionnaires, forming at total of 26 questions. Attention was paid so that a subject taking a particular version of the questionnaire was able to respond to all different questions without listening to the same excerpt two consecutive times, a fact that made the listening tests somewhat less unpleasant.

V. Evaluation results

In total 46 normal hearing subjects completed questionnaires with content from the concert and 40 with content from the football match. The preference scores for each

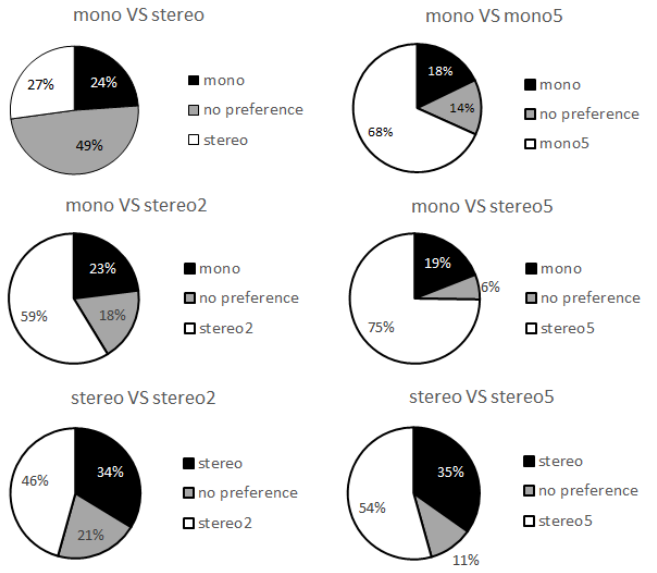

Fig. 3. Subjective preference in the case of the concert event (46 subjects).

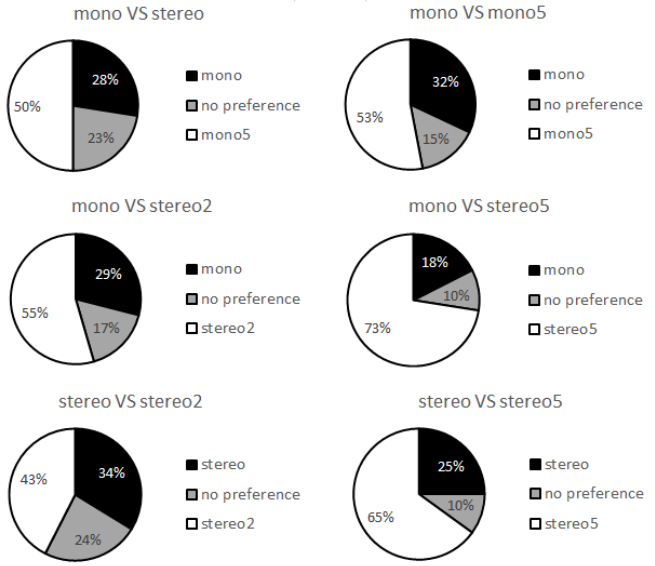

Fig. 4. Subjective preference in the case of the athletic event (40 subjects).

comparison case, obtained by grading responses as explained in section IV, are shown for the average of all responses in the athletic and the musical event in Fig. 2. In a more or less degree, it can be seen that the audio mixtures are preferred in comparison to the original UGRs, with some cases showing particularly great advantage. To asses if there is statistical difference between the two compared methods in each case, we transformed the subjects' responses so that they are compatible for a paired-reference test as follows. We assigned one vote to each method, regardless if a subject chose "better" or "slightly better" for stating his/her preference. Then, we equally distributed the " no preference" neutral responses to the two compared methods. Accounting for the number of trials devoted to each comparison case, we then calculated the minimum number of votes required towards one of the two methods in order to obtain statistical difference at various different levels of significance [14]. The results of the test indicate that for cases *mono* vs *mono5*, *mono* vs *stereo2*, *mono* vs *stereo5*, and *stereo* vs *stereo5*, statistical difference can be confirmed at $p = 0.01$ level of significance. On the other hand, pairs *mono* vs *stereo* and *stereo* vs *stereo2* do not pass the test at $p = 0.05$ level of significance.

For allowing a more in depth analysis of the subjects' responses, we also present pie charts with respect to each comparison case, for the concert in Fig. 3 and for the athletic event in Fig. 4. The pie charts illustrate how the subjects choices were distributed in each case, regardless if a subject chose "better" or "slightly better" for stating his or her preference.

Starting the discussion from case *mono* vs *mono5*, it can be seen that the listeners showed an important preference towards *mono5* in the case of the musical concert. In particular, *mono5* was judged to be better than *mono* in 68% of the cases, while the mono version was preferred against *mono5* only in 18% of the cases, as shown in the corresponding pie chart in Fig. 3. While in the majority of the cases *mono5* was preferred against *mono* also in the case of the athletic event, the relative advantage is here reduced to 53% versus 32%, to be discussed in mode detail in Section VI. Focusing on case *mono* vs *stereo2*, it can be seen that in both events there is a clear preference towards a stereo mixture of two UGRs compared to individual mono UGRs. Moreover, *stereo2* establishes and advantage compared to *mono* which is significantly higher than that obtained when comparing *stereo* to *mono*. To our opinion, a stereo signal resulting from the combination of two synchronized but possibly distant UGRs has less coherent left and right channels as opposed to when these two channels originate from the same device, and this seems to be positively appreciated by the subjects. It must be noted that during the test, the subjects were not informed about whether an audio stream is in mono or stereo format.

Finally, the comparisons *mono* vs *stereo5* and *stereo* vs *stereo5* demonstrate that the stereophonic mixture of five UGRs achieves the strongest advantage among all other mixing approaches. Indeed, *stereo5* is preferred in 64 % and 54% of the cases in comparison to *stereo* for the case of the athletic and the musical event respectively. The advantage in comparison to *mono* class is even more impressive, as *stereo5* is preferred in approximately 75% of the cases in both events.

VI. Discussion

The comparative listening tests indicated that even simple forms of content combination can transmit a significantly better listening experience, in comparison to individual UGRs, but to some degree, it would be desired to provide more specific reasons for this improvement. For simplifying the process, we asked subjects to provide their choices with respect to personal preference only, avoiding reference to more specific aspects

of quality. Still however, we can make some very reasonable assumptions about the reasons that led listeners to give higher preference scores to the mixed UGR classes.

Apart from the obvious advantage gained when comparing a monophonic to a stereophonic audio stream, we believe that content superposition acted simultaneously as a signal quality and a content quality enhancement process. To the authors opinion, the improvement in terms of signal quality resulted from the fact that, as parts of individual UGRs were possibly corrupted by distortion or poor frequency response, when multiple UGRs were added together, these problems were masked in the final mix. We refer the reader to the work of Fazenda et al in [15] for a discussion about regular quality problems in user contributed audio recordings. With respect to content quality, the improvement to some degree resulted from the fact that certain sound components were not captured at a particular recording location, but the same components were much more clearly heard in the audio stream produced from another location. Furthermore, superposition amplified the most interesting sound components - which were common across the different recordings - while at the same time, noise and interference - which was unique at each recording location - was de-emphasized in the final mix. Interestingly, by informal discussions with some of the subjects, we realized that the latter was judged to be an advantage more in the case of the concert than in the case of the football match, where speech or applause from individual spectators sitting close to the recording devices was considered by some to belong to the interesting content of the recording. While this somehow explains why *mono5* was preferred against *mono* in a larger degree in the concert than in the athletic event, we also believe that superposition of multiple distant UGRs in the athletic event resulted to speech from individual spectators becoming less audible. We invite the interested reader to listen to examples from both the original and mixed UGR classes, provided online in [16], in order to verify these arguments for himself/herself.

VII. Conclusion

Having a collection of synchronized UGRs as a basis, we formulated a monophonic and a stereophonic approach for mixing the available content. Subjective tests based on data from two different real-life public events showed that even with very simple mixing approaches and even with only five UGRs, an improved representation of the acoustic event can be constructed and delivered to the user. Certainly, there is a lot of space for further improvement, by making a more careful selection of the mixing parameters or by employing more advanced signal processing schemes. For example, in this evaluation, the selected panning weights were randomly assigned on the available UGRs. In a more advanced setting, these weights can be possibly defined based on the spatial coordinates of each recording device, or based on an estimation of the proximity between UGRs [17]. Also, an extension of the stereophonic mixing approach to a reproduction system with a greater number of channels can be straightforwardly conceptualized using Vector Based Amplitude Panning [18]. Finally, this investigation should also be extended to cases where the number of UGRs participating in the mix varies in the studied time interval. This is a more challenging case as time-varying weights would be required in order to avoid unwanted transitions in the perceived signal level or in the transmitted spatial impression.

VIII. Acknowledgement

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687605, Project COGNITUS.

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. J. 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] 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.

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

[10] N. Stefanakis, S. Chonianakis, and A. Mouchtaris, "Two open access datasets of user generated audio recordings," 2016. [Online]. Available: https://doi.org/10.5281/zenodo.167311

[11] ——, "Two datasets with user generated audio recordings," 2016. [Online]. Available: https://doi.org/10.5281/zenodo.164175

[12] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *International Society for Music Information Retrieval (ISMIR)*, 2002, pp. 107–115.

[13] "Loudness toolbox." [Online]. Available: http://genesis-acoustics.com/index.php?page=32

[14] E. Roessler, R. Pangborn, J. Sidel, and H. Stone, "Expanded statistical tables for estimating significance in paired-preference, paired-difference, duo-trio and triangle tests," *Journal of Food Science*, vol. 43, no. 3, pp. 940–943, 1978.

[15] B. Fazenda, P. Kendrick, T. Cox, F. Li, and I. Jackson, "Perception and automated assessment of audio quality in user generated content," in *Audio Engineering Society Convention 139*, Oct 2015.

[16] "Audio collection." [Online]. Available: http://users.ics.forth.gr/nstefana/eusipco17

[17] D. Ellis, H. Satoh, and Z. Chen, "Detecting proximity from personal audio recordings," in *Proc. of INTERSPEECH*, 2014, pp. 2519–2523.

[18] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.