Kernel Method for Speech Source Activity Detection in Multi-modal Signals

David Dov, Ronen Talmon and Israel Cohen
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel

***Abstract*—We consider a problem setup, in which a desired speech source is measured by a microphone and by a video camera in an interfering environment. We assume that the interfering sources in the audio signal are independent of the interfering sources in the video signal (e.g., the video signal does not capture the interfering speakers). Our objective in this paper is to detect the activity of the desired source. To address this problem, we take a kernel based geometric approach for obtaining a representation of the measured signal, in which the effect of the interfering sources is reduced. Based on this representation, we devise a measure for the activity of the desired source; experimental results demonstrate its superiority compared to competing methods in the detection of speech signals in the presence of different challenging types of interferences, including interfering speakers in the audio signal.**

***Index Terms*—Multi-modal signal processing, kernel methods, audio-visual speech activity detection.**

I. Introduction

We address the problem of activity detection of a speech source, measured both by a microphone and by a video camera pointed at the face of the speaker. We term this source as "the desired source". Assuming that it is measured in the presence of interferences, the objective in this paper is to detect the desired source while ignoring the interferences. We consider different types of interfering sources (interferences) in the audio signal, such as speech from other speakers, environmental noises, and transients, which are abrupt interruptions such as door-knocks [1]–[3]. The video signal may contain interferences such as head and mouth movements, which make the detection of the desired source difficult. Our main assumption is that the interferences in the two modalities (audio and video) are independent of each other, e.g., the video camera does not capture the interfering speakers.

The activity detection of the desired speech source may be useful for a variety of applications such as speech enhancement, speech and speaker recognition and speech diarization, where the goal is to determine "who spoke when" [4]–[8]. Speech diarization, for example, is a challenging problem since first, time intervals with active speech have to be accurately detected while ignoring both background noises, and transients, which often appear similar to speech [9], and second, the different speakers have to be distinguished, typically by assuming statistical models. In the audio-visual setting considered here, the activity of the desired source directly implies that the corresponding speaker is speaking regardless of presence or absence of interfering sources.

To address the problem of desired source activity detection, we take a multi-modal geometric approach, where the goal is to learn a representation of the data by exploiting relations (affinities) between data points in the different modalities (audio and video). Classical kernel based geometric methods, e.g., those presented in [10]–[14], typically address the problem of non-linear dimensionality reduction of single-modal data. They are based on constructing an affinity kernel capturing relations between the data points, and provide a low dimensional representation via the eigenvalue decomposition of the affinity kernel. Recent studies suggest to extend these kernel based geometric methods to the multi-modal case by constructing separate affinity kernels for each modality, and then by fusing the modalities through different combinations of the affinity kernels, e.g., by their weighted sum [15]–[27].

Lederman and Talmon presented in [26] a multi-modal fusion approach, where the data in the different modalities is fused by a product of affinity kernels, constructed separately for each modality. This fusion approach is particularly useful for the representation of the desired audio-visual source since, according to the analysis presented in [26], it reduces the effect of modality-specific sources, which in our problem setting are the interferences, by assumption. Hence, the obtained representation respects the relations between the data points according to the source present in both the modalities, which is the desired source in our case; therefore, it is particularly suitable for the activity detection of the desired source. In [28], we analyzed this fusion approach in a discrete setting showing that it may be further improved by a proper selection of the kernel bandwidth.

In this paper, we propose an algorithm for activity detection of a desired speech source. The algorithm is based on constructing two affinity kernels, one for each modality (audio and video), in a domain of features, separately built for each modality. We fuse the modalities by a product of the affinity kernels as in [26], [28] and devise a measure for the presence of the desired source using the eigenvalue decomposition of the product kernel. We apply the proposed algorithm for the detection of audio-visual speech signals in

This Research was supported by Qualcomm Research Fund and MAFAAT-Israel Ministry of Defense.

978-1-5090-2152-9/16/$31.00

the presence of multiple interfering audio sources including different speakers, background noises, and transients. Our simulation results demonstrate improved detection scores compared to single-modal variants, which are based on either the audio or the video signals, as well as compared to alternative fusion schemes.

We note that we consider as the main challenge in this study, the presence of *multiple* interfering sources. Specifically, we consider interferences that are of the same type as the desired source, i.e., other speakers in the audio signal. In addition, the video signal comprises the entire face of the speaker; therefore, head movements are considered interferences in the video. We note that in [28], we addressed a special case of the problem that is considered here; previously, we considered the presence of only a single interfering transient noise source, which is considerably different from speech. In addition and in contrast to this paper, only the mouth region of the speaker was assumed as the video signal, requiring an accurate detection of the mouth region as a preprocessing stage.

II. Problem Formulation

Consider a speech signal measured by a single microphone and by a video camera pointed at the face of a speaker. The signal is processed in consecutive frames, which are assumed aligned; let $\mathbf{v}_n \in \mathbb{R}^{L_v}$ and $\mathbf{w}_n \in \mathbb{R}^{L_w}$ be feature representations of the nth time frame in the first and the second modalities (i.e., audio and video), respectively, such that L_v and L_w are the total number of features in each modality. We use the Mel-Frequency Cepstral Coefficients (MFCC) [29] and motion vectors [30], for the representation of the audio and the video signals, respectively, as we describe in detail in [3]. The MFCCs are widely used for the representation of audio signals, and the motion vectors capture the movement of the mouth within the video, assumed to be associated with speech. In both modalities, we aggregate the features of three consecutive frames such that $\mathbf{v}_n$ is given by the MFCCs of frames $n-1, n, n+1$. Consider a sequence of N such pairs of frames:

$$\{(\mathbf{v}_n, \mathbf{w}_n)\}_{n=1}^{N}. \tag{1}$$

We assume that the measured audio signal comprises $M^v + 1$ sources: $S_1, S_2, ...S_{M^v}$ and S^d, where the superscript d stands for the desired source. Namely, the audio frame $\mathbf{v}_n$ is given by a mapping, denoted by f, of the sources to the features space:

$$\mathbf{v}_n = f(S^d, S_1, S_2, ...S_{M^v}).$$

The video signal comprises the video recording of the face of a speaker. Yet, there may be both natural mouth and head movements, which are not directly related to speech and are considered as interferences. Assuming M^w such interfering sources, the corresponding video frame $\mathbf{w}_n$ is given by:

$$\mathbf{w}_n = g(S^d, S_1, S_2, ...S_{M^w}),$$

where g denotes the mapping of the sources to the feature space of the video signal. With the exception of the desired source, the sources of the audio and the video signals are assumed independent. In addition, the sources are assumed to be present or absent independently of each other. Specifically, we assume two hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, for the absence and the presence of the desired source, respectively. Accordingly, let $\mathbb{1}_n$ be an indicator for the presence of the desired source in the nth frame, given by:

$$\mathbb{1}_n = \left\{ \begin{array}{ccc} 1 & ; & n \in \mathcal{H}_1 \\ 0 & ; & n \in \mathcal{H}_0 \end{array} \right\}. \tag{2}$$

The goal in this study is to detect the activity of the desired source, i.e., to estimate the indicator in (2).

III. Desired Speech Source Activity Detection

A. Multi-modal Fusion via the Product of Affinity Kernels

For completeness, we describe the fusion process based on a product between affinity kernels constructed separately for each modality, as proposed in [26]. Let $\mathbf{K}_v \in \mathbb{R}^{N \times N}$ be an affinity kernel of the first modality (i.e., audio), whose (n,m)th entry, denoted by $K_v(n,m)$, is given by:

$$K_v(n,m) = \exp\left(-\frac{||\mathbf{v}_n - \mathbf{v}_m||^2}{\epsilon_v}\right), \tag{3}$$

where ϵ_v is the kernel bandwidth whose selection we studied in [28]. By dividing each column by its sum, we construct a row stochastic matrix, which is denoted by $\mathbf{M}_v \in \mathbb{R}^{N \times N}$, and its (n,m)th entry, $M_v(m,n)$, is given by:

$$M_v(n,m) = \frac{K_v(n,m)}{d_v(n)}, \tag{4}$$

where $d_v(n) = \sum_{m=1}^{N} K_v(n,m)$. Similarly to $\mathbf{M}_v$, we construct a row stochastic matrix $\mathbf{M}_w \in \mathbb{R}^{N \times N}$ for the second modality, and the data from the two modalities are fused by the product of the row stochastic matrices:

$$\mathbf{M} = \mathbf{M}_v \cdot \mathbf{M}_w, \tag{5}$$

where $\mathbf{M} \in \mathbb{R}^{N \times N}$ is viewed as aggregating the relations between the data points in the two modalities. Lederman and Talmon considered in [26] the continuous counterparts of $M_v(n,m)$, $M_w(n,m)$ and $M(n,m)$ as diffusion operators. They showed that the continuous operator corresponding to $M(n,m)$ is an alternating diffusion operator, which integrates out modality-specific sources by applying the diffusion process in two steps corresponding to the two modalities.

B. Desired Source Activity Detection

For the detection of the desired source, we apply an eigenvalue decomposition to $\mathbf{M}$. The eigenvectors respect the relations between the multi-modal data points aggregated in the matrix $\mathbf{M}$, and therefore, they are often used in the literature to form a low dimensional representation of the data [10]. The matrix $\mathbf{M}$ is row stochastic since $\mathbf{M}_v$ and $\mathbf{M}_w$ are row stochastic matrices, so it has an all ones eigenvector corresponding to the eigenvalue one, which we neglect since it does not contain information [10]. Since $\mathbf{M}$ integrates out

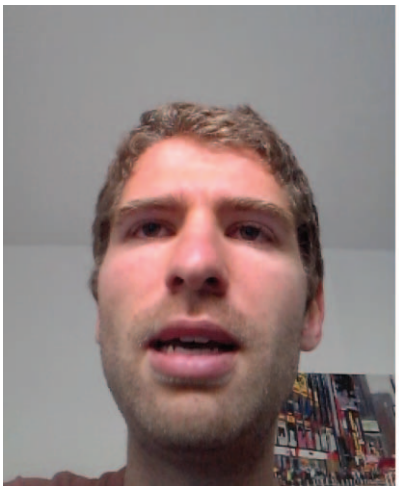


Fig. 1: An example of a video frame.

the modality-specific sources, which are the interferences in our case, its eigenvectors represent the data according to the desired audio-visual source. For the detection of the desired source, we use the leading (non-trivial) eigenvector, which we denote by $\boldsymbol{\nu}_1 \in \mathbb{R}^N$; the nth entry of $\boldsymbol{\nu}_1$, denoted by $\nu_1(n)$, corresponds to the nth frame of the measured signals (audio and video) and we view it as a new mapping h of the nth frame according to the desired source:

$$\nu_1(n) = h(S^d).$$

The leading eigenvector of an affinity kernel is typically used in the literature for clustering such that the nth data point is clustered according to the sign of $\nu_1(n)$ [31]. Indeed, we have found in our experiments, that the data are properly clustered by $\boldsymbol{\nu}_1$ according to the presence and the absence of the desired source. Accordingly, we propose to estimate the indicator for the presence of the desired source $\mathbb{1}_n$ in (2) by comparing the eigenvector entries to a threshold τ:

$$\hat{\mathbb{1}}_n = \left\{ \begin{array}{ccc} 1 & ; & \nu_1(n) > \tau \\ 0 & ; & \text{otherwise} \end{array} \right\}. \tag{6}$$

Namely, we view the leading eigenvector as a continuous measure of the presence of the desired source. The threshold τ controls the trade-off between the probability of correct detection of the desired source and the probability of false alarm, and its setting is application dependent.

IV. Experimental Results

We consider an audio-visual recording of a speaker measured by a microphone and by a video camera pointed at the face of the speaker. We use a dataset, which we presented in [3], comprising 11 sequences of different speakers, 60 s long each. The video signal is measured in 25 fps frame rate, and the audio signal, which is measured in 8 kHz, is aligned to the video signal using frames of 634 samples with 50% overlap. To simulate the interferences, we synthetically add to the audio signal different types of background noises and transients taken from a free online corpus [32], and other speakers taken from the dataset in [3]. The video signal comprises the entire face of the speaker as demonstrated in Fig. 1, in contrast to [28], where cropping of the mouth region of the speaker was required as a preprocessing step. Therefore, it may contain natural head and mouth movements, which are not related to speech. To set the ground truth of the activity of the desired speech source (which also appears in the video), we use the clean audio signal and consider the desired source active in a frame if its energy level is above 1% from the maximal energy value in the sequence. In this type of ground truth setting, the resolution of the presence and absence of the desired source is up to a single frame and it may be used, for example, for the enhancement of the desired source [8].

An example of the detection of the desired speech source obtained by the proposed algorithm is presented in Fig. 2, where we consider three audio sources comprising two speakers – one desired, one interfering and babble noise. For the clarity of presentation, we use a relatively high signal to noise ratio (SNR) of 20 dB, where the SNR is calculated with respect to the desired speaker and the babble noise. Hence, the main challenge in this example is to distinguish between the desired speech signal and the speech signal of the interfering speaker. Indeed, the spectrogram of the measured audio signal, presented in Fig. 2 (Bottom), demonstrates that just by observing the spectrogram, it is hard to distinguish between the speech parts corresponding to the desired speech and the interfering speech. In Fig. 2 (Top), we qualitatively compare the proposed method for the detection of the desired source to an alternative kernel method termed "Hadamard", in which, instead of the product between the kernels in (5), the modalities are fused by the Hadamard product: $\mathbf{M}_v \circ \mathbf{M}_w$, where $\circ$ denotes pointwise multiplication. For both approaches, we set the value of the threshold τ in (6) to provide 80% correct detection rate and compare their false alarm rates. It may be seen that the proposed approach provides significantly fewer false alarms, and the competing method wrongly detects the activity of the interfering speech source, e.g., in the time interval after the 24th second.

In Fig. 3 we present the results of a quantitative evaluation of the proposed approach in the form of receiver operating characteristic (ROC) curves, which are plots of detection versus false alarm rates. The proposed approach is compared, in addition to the method "Hadamard", to a method based on fusing the modalities via a sum of the affinity kernels, i.e., $\mathbf{M}_v + \mathbf{M}_w$, termed "Sum" in the plots. In addition, we compare the proposed approach to its single-modal variants, termed "Audio" and "Video", which are based on estimating the speech indicator in (6) using the leading eigenvector of the kernels $\mathbf{M}_v$ and $\mathbf{M}_w$, respectively. We observe in the plots that the approaches based on a single modality attain comparable results; the detector of the desired source based only on the audio signal is limited due to high similarity of the desired

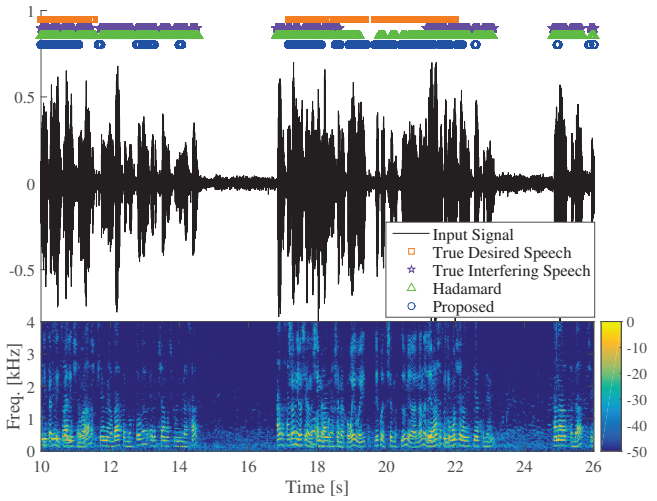
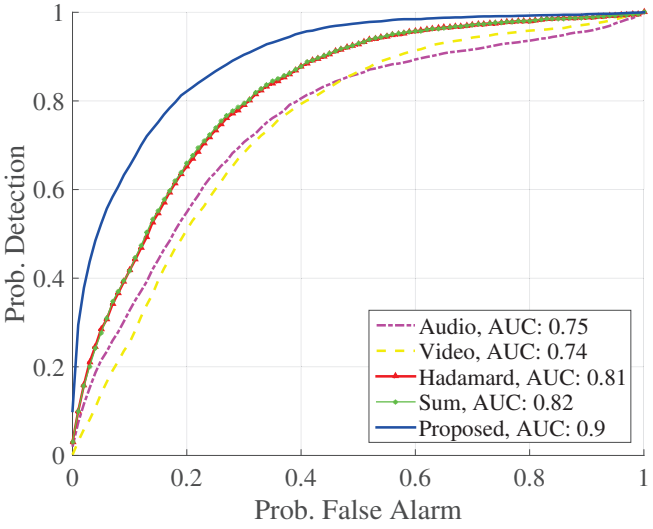
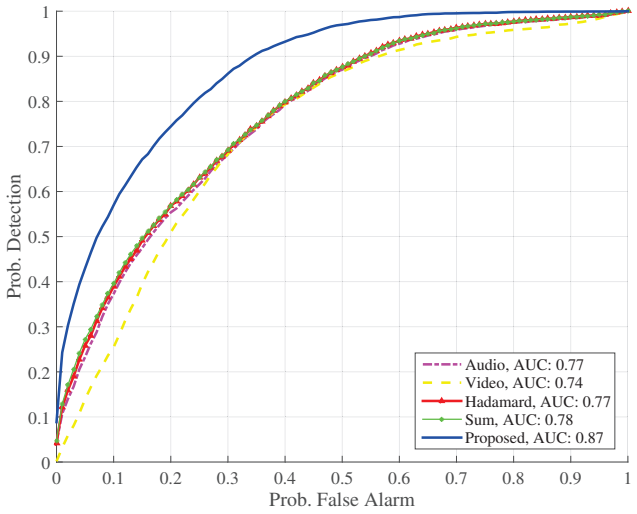


Fig. 2: Qualitative assessment of the proposed algorithm for the desired speech source activity detection in the presence of three sources: the desired speech source, an interfering speech source, and babble noise with 20 dB SNR. (Top) Time domain, input signal - black solid line, true desired speech source - orange squares, true interfering speech source - purple stars, "Hadamard" with a threshold set for 80% correct detection rate - green triangles, proposed algorithm with a threshold set for 80% correct detection rate- blue circles. (Bottom) Spectrogram of the input signal.

source especially to the other speakers. The performance based only on the video signal are also limited both due to modality-specific sources such as movements of the head and due to the high resolution of the ground truth. Indeed, there exist speech parts that do not involve the movement of mouth in certain time frames. The alternative fusion schemes perform slightly better than the single-modal approaches. Finally, the proposed approach for the detection of the desired source outperforms all other methods and provides improved detection scores for all false alarm values.



(a)



(b)

Fig. 3: Probability of the detection vs probability of false alarm. Source types: (a) two speakers and babble noise with 20 dB SNR, (b) two speakers, door-knocks transients and white Gaussian noise with 15 dB SNR.

V. Conclusion

We have addressed the problem of audio-visual speech source detection in the presence of interferences. We proposed an algorithm for the detection of a desired source by fusing the modalities via a product of kernels, constructed separately for each modality. An eigenvalue decomposition of the product kernel yields a useful representation of the data, in which the effects of the interfering sources are reduced, allowing us to devise a measure of the presence of the desired source based on the leading eigenvector. Experimental results have demonstrated the improved performance of the proposed algorithm in challenging environments, including speech activity detection in audio-visual data under presence of modality-specific interfering sources.

References

[1] A. Hirszhorn, D. Dov, R. Talmon, and I. Cohen, "Transient interference suppression in speech signals based on the OM-LSA algorithm," in *Proc. International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012, pp. 1–4.

[2] D. Dov and I. Cohen, "Voice activity detection in presence of transients using the scattering transform," in *IEEE 28th Convention of Electrical and Electronics Engineers in Israel (IEEEI), 2014*, Dec 2014, pp. 1–5.

[3] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, April 2015.

[4] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[5] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, 2002.

[8] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[9] D. Dov, R. Talmon, and I. Cohen, "Kernel method for voice activity detection in the presence of transients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1–1, 2016.

[10] R.R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[11] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. the National Academy of Sciences*, vol. 100, no. 10, pp. 5591–5596, 2003.

[12] M.l Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[13] M. Balasubramanian, E. L. Schwartz, Tenenbaum J. B., de Silva V., and J. C. Langford, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[14] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[15] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1159–1166.

[16] M. B. Blaschko and C. H. Lampert, "Correlational spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.* IEEE, 2008, pp. 1–8.

[17] V. R. De Sa, P. W. Gallagher, J. M. Lewis, and V. L. Malave, "Multi-view kernel construction," *Machine learning*, vol. 79, no. 1-2, pp. 47–71, 2010.

[18] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 1413–1421.

[19] A. Kumar and H. Daumé, "A co-training approach for multi-view spectral clustering," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 393–400.

[20] Y. Y. Lin, T. L. Liu, and C. S0 Fuh, "Multiple kernel learning for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 6, pp. 1147–1160, 2011.

[21] B. Wang, J. Jiang, W. Wang, Z. H. Zhou, and Z. Tu, "Unsupervised metric fusion by cross diffusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. IEEE, 2012, pp. 2997–3004.

[22] H. C. Huang, Y. Y. Chuang, and C. S. Chen, "Affinity aggregation for spectral clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012*. IEEE, 2012, pp. 773–780.

[23] B. Boots and G. Gordon, "Two-manifold problems with applications to nonlinear system identification," *arXiv preprint arXiv:1206.4648*, 2012.

[24] O. Lindenbaum, A. Yeredor, M. Salhov, and A. Averbuch, "Multiview diffusion maps," *arXiv preprint arXiv:1508.05550*, 2015.

[25] O. Lindenbaum, A. Yeredor, and M. Salhov, "Learning coupled embedding using multiview diffusion maps," in *Latent Variable Analysis and Signal Separation*, pp. 127–134. Springer, 2015.

[26] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Applied and Computational Harmonic Analysis*, 2015.

[27] T. Michaeli, W. Wang, and T. Livescu, "Nonparametric canonical correlation analysis," *Submitted to International Conference on Learning Representations (ICLR 2016)*.

[28] D. Dov, R. Talmon, and I. Cohen, "Kernel-based sensor fusion with application to audio-visual voice activity detection," *arXiv preprint arXiv:1604.02946*, 2016.

[29] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. 1st International Conference on Music Information Retrieval (ISMIR)*, 2000.

[30] A. Bruhn, J. Weickert, and C. Schnörr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods," *International Journal of Computer Vision*, vol. 61, no. 3, pp. 211–231, 2005.

[31] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[32] *[Online]. Available: http://www.freesound.org*.