MAPPING SOUNDS ONTO IMAGES USING BINAURAL SPECTROGRAMS

Antoine Deleforge[1], *Vincent Drouard*[1], *Laurent Girin*[1,2,3], *and Radu Horaud*[1]

[1]INRIA Grenoble Rhône-Alpes [2]GIPSA-Lab [3]Univ. Grenoble Alpes, France

ABSTRACT

We propose a novel method for mapping sound spectrograms onto images and thus enabling alignment between auditory and visual features for subsequent multimodal processing. We suggest a supervised learning approach to this audio-visual fusion problem, on the following grounds. Firstly, we use a Gaussian mixture of locally-linear regressions to learn a mapping from image locations to binaural spectrograms. Secondly, we derive a closed-form expression for the conditional posterior probability of an image location, given both an observed spectrogram, emitted from an unknown source direction, and the mapping parameters that were previously learnt. Prominently, the proposed method is able to deal with completely different spectrograms for training and for alignment. While fixed-length wide-spectrum sounds are used for learning, thus fully and robustly estimating the regression, variable-length sparse-spectrum sounds, e.g., speech, are used for alignment. The proposed method successfully extracts the image location of speech utterances in realistic reverberant-room scenarios.

1. INTRODUCTION

The association of auditory and visual data has been proved to considerably improve the robustness of speech processing whenever close-distance and frontal recordings of the speaker's face or lips are available, for example to separate speech from noise [1–3], from competing sources [4, 5], or for speech recognition [6–8]. However, the performance of these methods rapidly degrades with far distance (2-3 meters) cameras and microphones. The problem becomes more difficult for two reasons: several objects, e.g., faces, are present at once, and both the visual and auditory data are degraded. Vision is useful, among others, for detecting faces and facial behaviors. Nevertheless, the field of view and the range of a camera are limited and the image features are perturbed by occlusions, perspective effects, illumination, and the relative position between camera and scene. In contrast, auditory signals have the potential to enable verbal interactions but they are spatially ubiquitous, subject to overlap, and corrupted by the acoustic environment.

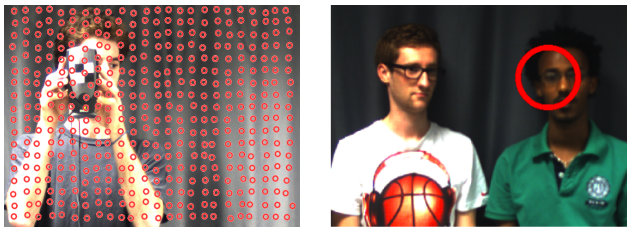

Fig. 1. *Left*: The sound-to-image transformation is learned from an audio-visual target composed of a loud speaker and a visual marker. The loud-speaker emits full-spectrum sounds which are recorded with a microphone pair while the visual marker is used to estimate its corresponding image positions. *Right*: A speaking face is detected and predicted from a sparse-spectrum sound.

Hence, vision and audition have both strengths and limitations and exploiting their complementarity is still a challenging problem. In this paper we propose a method that aligns auditory and visual data based on *mapping auditory spectrograms onto images*. We start by learning a sound-to-image transformation from a training set of audio/visual pairs (Fig. 1-left); this mapping is then applied to a sound source that emits from an unknown image location (Fig. 1-right). Several problems need to be addressed for such a method to be effective: (i) the sound representation should encode its location with respect to the microphones while it should be independent of the spectral content of the sound, (ii) in order to properly and fully estimate all the sound-to-image mapping parameters, the training should be performed with full-spectrum signals, and (iii) natural sounds such as speech signals, which are of particular interest, have sparse spectrograms and it is not clear how to map them onto images, based on the sound-to-image transformation parameters that were estimated with full-spectrum sounds.

We propose to use a setup composed of one camera and two microphones. The microphones are embedded into an acoustic dummy head, yielding non-linear filtering effects that depend on the sound direction (azimuth and elevation). We stress that with such a setup one can build *interaural spectrograms* that encode *two-dimensional* (2D) directional information and which are independent of the spectral con-

Support from the European Research Council (ERC) through the Advanced Grant VHIA (#340113) is greatly acknowledged.

tent of the emitted sound [9]. In order to estimate a mapping from sounds to images, we propose to learn the parameters of a Gaussian mixture of locally linear regression functions using pairs of *full-spectrum-data* and *pixel coordinates of associated image landmarks*.

However, estimating such a function *directly*, from high-dimensional spectrograms to low-dimensional image locations, is problematic for two reasons. Firstly, the large number of parameters that need to be estimated in this case is prohibitive. Secondly, as detailed below, it is not possible to apply this direct function to natural sounds. Instead, we train an *image-to-audio* regression (Section 2) using full-spectrum (white noise) signals. Then, we derive an analytic expression for the conditional posterior distribution of an image location, given a sparse spectrogram and the learned regression parameters (Section 3). We show that this posterior distribution is a Gaussian mixture whose parameters (priors, means, and covariances) can be computed in closed-form, from the parameters of the regression function. Prominently, the proposed solution can deal with missing data in the observed acoustic vectors, for example, it is able to map natural sounds, such as speech, onto images of faces.

The proposed method is strongly linked with the problem of sound-source localization. This is traditionally done using time difference of arrival (TDOA) or intensity-level differences (ILD) between microphones, e.g., [10–12] to cite just a few. However these methods perform only one-dimensional (azimuth) source localization. In the framework of audio-visual alignment, 2D estimation is needed in order to associate sound-source directions with image locations, which rules out azimuth-only source localization. Another possibility is to estimate the geometry of the camera-microphone setup, but this turns out to be difficult because of the non-linear nature of the sound-propagation model [13]. TDOA-based 2D sound-source localization needs non-coplanar microphone arrays and the solution must solve a complex non-linear constrained optimization problem [14].

In [9] it was experimentally shown that the space of interaural spectral features is homeomorphic to a 2D manifold that can be parameterized by azimuth and elevation. A dummy head mounted onto a pan-tilt mechanism was used to collect pairs of motor positions and interaural spectrograms, used to train a regression function. However, the emitter was kept static at a single position in all training and test experiments, while the dummy head was rotated onto itself. The method was hence limited to theoretical conclusions rather than practical applications. In this paper, we extend [9] by formally stating and proving a theorem showing that natural sound spectrograms can be mapped onto an image plane by inverting a piecewise affine mapping. We subsequently introduce a novel and elegant way of precisely locating a sound source, based on its pixel location, without moving the sensors.

2. LEARNING WITH FULL SPECTRUM SOUNDS

Let us consider a single, *full-spectrum* (*e.g.* white noise) point source, and a static auditory system that captures acoustic vectors in $\mathcal{Y} \subset \mathbb{R}^D$ along time. We assume that at any time, these vectors depend on the source position but not on its emitted spectrum. Let $\mathcal{X} \subset \mathbb{R}^2$ be a set of image locations, asociated to a static camera field of view. We consider N training pairs $\{(\boldsymbol{y}_n, \boldsymbol{x}_n)\}_{n=1}^{n=N} \subset \mathcal{Y} \times \mathcal{X}$ that associate acoustic vectors with their corresponding sound source position in the image. These pairs are realizations of observed random variables $(\boldsymbol{Y}, \boldsymbol{X})$. We consider the following piecewise linear regression from the low-dimensional image plane to the high-dimensional acoustic space:

$$\boldsymbol{Y} = \sum_{k=1}^{K} \mathbb{I}(Z = k)(\mathbf{A}_k \boldsymbol{X} + \boldsymbol{b}_k) + \boldsymbol{E} \tag{1}$$

where matrix $\mathbf{A}_k \in \mathbb{R}^{D \times 2}$ and vector $\boldsymbol{b}_k \in \mathbb{R}^D$ are the parameters of an affine transformation τ_k and $\boldsymbol{E} \in \mathbb{R}^D$ is a Gaussian error vector with zero-mean and diagonal covariance $\boldsymbol{\Sigma} = \text{Diag}\,(\sigma_1^2 \dots \sigma_d^2 \dots \sigma_D^2)$ capturing both the observation noise in $\mathbb{R}^D$ and the reconstruction error due to the local affine approximation. $\mathbb{I}$ is the indicator function and Z is a hidden variable such that $\mathbb{I}(Z = k) = 1$ if and only if $Z = k$ ($\boldsymbol{Y}$ is the transformed of $\boldsymbol{X}$ by τ_k), and 0 otherwise. Consequently we have $p(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}, Z = k) = \mathcal{N}(\boldsymbol{y}; \mathbf{A}_k \boldsymbol{x} + \boldsymbol{b}_k, \boldsymbol{\Sigma})$. To constrain the affine transformations to be local, we associate the K affine transformations to an equal number of regions $\{\mathcal{R}_k\}_{k=1}^{k=K} \subset \mathbb{R}^2$ that define a partitioning of $\mathcal{X}$. The regions are modeled in a probabilistic way by assuming that $\boldsymbol{X}$ follows a mixture of K Gaussians defined by $p(\boldsymbol{X} = \boldsymbol{x} | Z = k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ with prior $p(Z = k) = \pi_k$ and with $\boldsymbol{c}_k \in \mathbb{R}^2$, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{2 \times 2}$, and $\sum_{k=1}^{K} \pi_k = 1$. To summarize, the model parameters are

$$\boldsymbol{\theta} = \{\{\boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \pi_k, \mathbf{A}_k, \boldsymbol{b}_k\}_{k=1}^{K}, \boldsymbol{\Sigma}\} \tag{2}$$

and they can be estimated via an EM procedure yielding closed-form expressions for the model parameters [15].

3. MAPPING NATURAL SOUNDS ON IMAGES

We now consider the localization of natural *sparse-spectrum* sounds, e.g., speech. A sound is described by T acoustic vectors forming a time series, namely $\mathbf{Y}' = \{\boldsymbol{y}'_1 \dots \boldsymbol{y}'_t \dots \boldsymbol{y}'_T\} \subset \mathbb{R}^D$. We assume that these acoustic vectors are emitted from the same location. A time series $\mathbf{Y}'$ can be viewed as a $D \times T$ spectrogram and each entry y'_{dt} is referred to as a frequency-time point. Natural sounds are represented by spectrograms that are extremely sparse, i.e., many frequency-time points are null, resulting in an unusable or *missing* acoustic value at

that point. To account for this, we introduce a $D \times T$ matrix $\boldsymbol{\chi} = \{\chi_{dt}\}_{d,t=1}^{D,T}$ of binary variables such that $\chi_{dt} = 1$ if the frequency-time point is active and $\chi_{dt} = 0$ otherwise. To summarize, a test sound is described by $\mathcal{S} = \{\mathbf{Y}', \boldsymbol{\chi}\}$ and we seek the posterior density of the sound's image location, $p(\boldsymbol{x}|\mathcal{S}; \widetilde{\boldsymbol{\theta}})$. We state and proof a theorem allowing for the full characterization of this density.

Theorem 1 *Under the assumption that all the acoustic vectors in $\mathcal{S}$ are emitted from the same location, the posterior distribution is a Gaussian mixture model in $\mathbb{R}^2$, namely*

$$p(\boldsymbol{x}|\mathcal{S}; \widetilde{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \nu_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{V}_k). \tag{3}$$

whose parameters $\{\nu_k, \boldsymbol{\mu}_k, \mathbf{V}_k\}_{k=1}^{k=K}$ can be expressed in closed-form with respect to $\widetilde{\boldsymbol{\theta}}$ and $\mathcal{S}$, namely:

$$\boldsymbol{\mu}_k = \mathbf{V}_k \left(\widetilde{\boldsymbol{\Gamma}}_k^{-1} \widetilde{\boldsymbol{c}}_k + \sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2} \widetilde{\boldsymbol{a}}_{dk} (y'_{dt} - \widetilde{b}_{dk}) \right) \tag{4}$$

$$\mathbf{V}_k = \left(\widetilde{\boldsymbol{\Gamma}}_k^{-1} + \sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2} \widetilde{\boldsymbol{a}}_{dk} \widetilde{\boldsymbol{a}}_{dk}^\top \right)^{-1} \tag{5}$$

$$\nu_k \propto \widetilde{\pi}_k \frac{|\mathbf{V}_k|^{\frac{1}{2}}}{|\widetilde{\boldsymbol{\Gamma}}_k|^{\frac{1}{2}}} \exp\Bigg(-\frac{1}{2} \Big(\sum_{d,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2} (y'_{dt} - \widetilde{b}_{dk})^2 + \widetilde{\boldsymbol{c}}_k^\top \widetilde{\boldsymbol{\Gamma}}_k^{-1} \widetilde{\boldsymbol{c}}_k - \boldsymbol{\mu}_k^\top \mathbf{V}_k^{-1} \boldsymbol{\mu}_k \Big) \Bigg) \tag{6}$$

where $\widetilde{\boldsymbol{a}}_{dk}^\top \in \mathbb{R}^2$ is the d^{th} row of $\widetilde{\mathbf{A}}_k$, $\widetilde{b}_{dk} \in \mathbb{R}$ is the d^{th} entry of $\widetilde{\mathbf{b}}_k$ and ν_k is normalized to sum to 1 over k.

The posterior expectation can then be used to estimate the sound's image location: $\widehat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\mathcal{S}; \widetilde{\boldsymbol{\theta}}] = \sum_{k=1}^{K} \nu_k \boldsymbol{\mu}_k$.

Proof of theorem 1. By including the hidden variables Z (section 2) and using the sum rule, we obtain:

$$p(\boldsymbol{x}|\mathcal{S}; \widetilde{\boldsymbol{\theta}}) = \sum_{k=1}^{K} p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}}) p(Z = k|\mathcal{S}; \widetilde{\boldsymbol{\theta}}). \tag{7}$$

Since the proposed model implies an affine dependency between the Gaussian variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ given Z, the term $p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}})$ is a Gaussian distribution in $\boldsymbol{x}$. In other words, for each k, there is a mean $\boldsymbol{\mu}_k \in \mathbb{R}^2$ and a covariance matrix $\mathbf{V}_k \in \mathbb{R}^{2\times 2}$ such that $p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{V}_k)$. Notice that $\nu_k = p(Z = k|\mathcal{S}; \widetilde{\boldsymbol{\theta}})$ is not conditioned by $\boldsymbol{x}$. With these notations, (7) leads directly to (3). We now detail the computation of the GMM parameters $\{\boldsymbol{\mu}_k, \mathbf{V}_k, \nu_k\}_{k=1}^{K}$. Using Bayes inversion we have:

$$p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}}) = \frac{p(\mathcal{S}|\boldsymbol{x}, Z = k; \widetilde{\boldsymbol{\theta}}) p(\boldsymbol{x}|Z = k; \widetilde{\boldsymbol{\theta}})}{p(\mathcal{S}|Z = k; \widetilde{\boldsymbol{\theta}})}. \tag{8}$$

Since we already assumed that the measurement noise has a diagonal covariance, the observations in $\mathcal{S}$ are conditionally independent given Z and $\boldsymbol{x}$. Therefore, by omitting the denominator of (8) which does not depend on $\boldsymbol{x}$, it follows that $p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}})$ is proportional to

$$\begin{aligned} & p(\boldsymbol{x}|Z = k; \widetilde{\boldsymbol{\theta}}) \prod_{d=1,t=1}^{D,T} p(y'_{dt}|\boldsymbol{x}, Z = k; \widetilde{\boldsymbol{\theta}})^{\chi_{dt}} \\ & = \mathcal{N}(\boldsymbol{x}; \widetilde{\boldsymbol{c}}_k, \widetilde{\boldsymbol{\Gamma}}_k) \prod_{d=1,t=1}^{D,T} \mathcal{N}(y'_{dt}|\widetilde{\boldsymbol{a}}_{dk}^\top \boldsymbol{x} + \widetilde{b}_{dk}, \widetilde{\sigma}_d^2)^{\chi_{dt}} \\ & = \frac{C}{|\widetilde{\boldsymbol{\Gamma}}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (A + B) \right) \end{aligned} \tag{9}$$

where $A = \sum_{d=1,t=1}^{D,T} \frac{\chi_{dt}}{\widetilde{\sigma}_d^2} (y'_{dt} - \widetilde{\boldsymbol{a}}_{dk}^\top \boldsymbol{x} - \widetilde{b}_{dk})^2$, $B = (\boldsymbol{x} - \widetilde{\boldsymbol{c}}_k)^\top \widetilde{\boldsymbol{\Gamma}}_k^{-1} (\boldsymbol{x} - \widetilde{\boldsymbol{c}}_k)$, and the constant C depends neither on $\boldsymbol{x}$ nor on k. Since $p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}})$ is a normal distribution in $\boldsymbol{x}$ with mean $\boldsymbol{\mu}_k$ and covariance $\mathbf{V}_k$, we can write:

$$A + B = (\boldsymbol{x} - \boldsymbol{\mu}_k)^\top \mathbf{V}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k). \tag{10}$$

By identification between the left-hand and right-hand terms of (10), we obtain the formulae (4) and (5) for $\boldsymbol{\mu}_k$ and $\mathbf{V}_k$ respectively. Using Bayes inversion, the mixture's priors $\nu_k = p(Z = k|\mathcal{S}; \widetilde{\boldsymbol{\theta}})$ are proportional to $\widetilde{\pi}_k p(\mathcal{S}|Z = k; \widetilde{\boldsymbol{\theta}})$. Unfortunately, we cannot directly decompose $p(\mathcal{S}|Z = k; \widetilde{\boldsymbol{\theta}})$ into a product over (d, t), as previously done with $p(\mathcal{S}|\boldsymbol{x}, Z = k; \widetilde{\boldsymbol{\theta}})$. Indeed, while it is assumed that the frequency-time points of the observed spectrogram $\mathcal{S}$ are independent *given $\boldsymbol{x}$ and Z*, this is not true for the same observations *given only Z*. However, we can use (8) to obtain

$$p(\mathcal{S}|Z = k; \widetilde{\boldsymbol{\theta}}) = \frac{p(\mathcal{S}|\boldsymbol{x}, Z = k; \widetilde{\boldsymbol{\theta}}) p(\boldsymbol{x}|Z = k; \widetilde{\boldsymbol{\theta}})}{p(\boldsymbol{x}|\mathcal{S}, Z = k; \widetilde{\boldsymbol{\theta}})}. \tag{11}$$

The numerator is given by (9) and the denominator is the normal distribution $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \mathbf{V}_k)$. After simplifying the terms in $\boldsymbol{x}$, we obtain the desired expression (6) for ν_k. ■

4. EXPERIMENTS AND RESULTS

Setup. A binaural pair of microphones embedded into an acoustic dummy head is mounted onto a camera system[1]. The audio-visual source used for training is a loud-speaker with a visual marker, e.g., Fig. 1(left). We assume a one-to-one mapping between pixels and sound directions.

Interaural acoustic vectors. Let a sound source emit from a direction corresponding to a pixel position $\boldsymbol{x} \in \mathbb{R}^2$. We denote the complex-valued left- and right-spectrograms with $\mathbf{s}^{\mathrm{L}} = \{s_{ft}^{\mathrm{L}}\}_{t=1,f=1}^{F,T}$ and $\mathbf{s}^{\mathrm{R}} = \{s_{ft}^{\mathrm{R}}\}_{t=1,f=1}^{F,T}$. These spectrograms are built using the short-time Fourier transform with F frequency bands and T windows of size 64ms with 56ms overlap. The resulting spectrograms have 125 temporal bins

[1]Full details on the setup at http://team.inria.fr/perception/popeye/

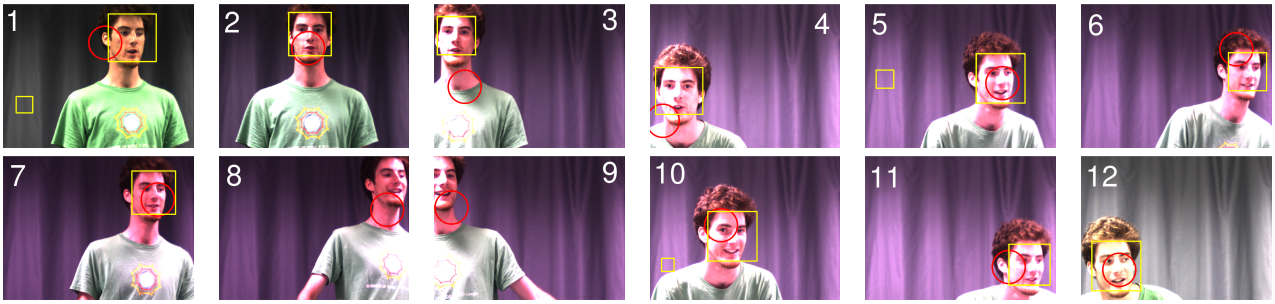

Fig. 2. A subject counts from 1 to 12 (white numbers) with a normal voice loudness and is static while pronouncing each number. The red circles show the position found with our method. The yellow squares show the results of the Viola-Jones face detector [16].

per second and 512 frequency bins ranging from 0 to 8KHz. The *interaural-level-difference* (ILD) and *interaural-phase-difference* (IPD) spectrograms, $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$, are defined as the log-amplitude and phase of the ratio between the left and the right spectrograms, *i.e.*, $\alpha_{ft} = 20\log_{10}(|s^{\mathrm{R}}_{ft}/s^{\mathrm{L}}_{ft}|) \in \mathbb{R}$ and $\phi_{ft} = \exp(j\arg(s^{\mathrm{R}}_{ft}/s^{\mathrm{L}}_{ft})) \in \mathbb{C} \equiv \mathbb{R}^2$. A binaural acoustic vector $\boldsymbol{y}$ is obtained by taking the temporal mean of the concatenation of $\boldsymbol{\alpha}$ and $\boldsymbol{\phi}$, hence we obtain a vector of size $D = 1536$. Because of the filtering effects from source to microphones, α_{ft} and ϕ_{ft} are independent of the emitted signal, and depend on the sound source position $\boldsymbol{x}$ if the recorded spectral density at the spectrogram point (f, t) is large enough [9]. Time-frequency points for which the total spectral density $10\log_{10}(|s^{\mathrm{L}}_{ft}|^2 + |s^{\mathrm{R}}_{ft}|^2)$ is below a given threshold are treated as missing values.

Training data[2] are obtained with a loudspeaker emitting white-noise (WN) from $N = 432$ different positions lying on a 18×24 grid in the camera's field of view, e.g., Fig. 1-left. Since a WN signal have a significant spectral density in all time-frequency points, the associated ILD and IPD spectrograms $\boldsymbol{\alpha}_n$ and $\boldsymbol{\phi}_n$ do not have missing values. Therefore we obtain a training set $(\boldsymbol{y}_n, \boldsymbol{x}_n), n \in [1 \dots N]$. In all our experiments we used $K = 32$ affine components to learn the regression (Section 2).

Loud-speaker test data[2] are obtained with the loud-speaker emitting 1-second utterances from the TIMIT dataset [17].The loud speaker is placed at 108 positions on a 9×12 grid covering the camera's field of view. The binaural recording associated to each utterance is cut, yielding ILD and IPD spectrograms with 89% missing values on average. These spectrograms are concatenated vertically, resulting in a time series of acoustic vectors $\mathcal{S} = \{\mathbf{Y}', \boldsymbol{\chi}\}$ (Section 3).

***Counting* test**[2]**.** The method was also tested on in a more realistic scenario. A participant is asked to count from 1 to 20 in front of the camera. The speaker is required to be roughly static while pronouncing each number, whereas he/she is allowed to move in between numbers. The audio-to-visual mapping method is applied on a 720ms analysis window that is slid over the soundtrack, in order to estimate a speaker position at each video frame for which enough acoustic level is recorded. This is a particularly challenging scenario because the speaker has different direction, distance, orientation and directionality than the loudspeaker used for training, and the speaker emits sparse-spectrum and less loud sounds than in the training set, thus considerably reducing the amount of exploitable spectrogram data.

Results. The average localization error over 108 loud-speaker tests was 21.9 ± 17 pixels horizontally and 23.1 ± 20 vertically. This corresponds to less that 1° in both azimuth and elevation. The largest *ground-truth-to-estimate distance* (GTED) error was 89.9 pixels, *i.e.* $\approx 4^\circ$. For comparison, we used PHAT [18] as a baseline sound source localization method, with the same test sounds. PHAT estimates the sound's *time difference of arrival* (TDOA), using cross-correlations at different times and frequency channels. A linear regressor was trained to map TDOA values obtained with PHAT onto the horizontal image axis using the WN training data[3]. Note that the vertical position cannot be estimated from TDOA. The average GTED with PHAT was 64.0 ± 51.5 pixels (3 times larger than with the proposed method) with 21 out of 108 GTED larger than 100 pixels. The average computational time of our method and of PHAT are of respectively 230ms and 400ms for 1s sounds.

Figure 2 shows some frames of the video generated from the counting test[2]. The sound source position estimated by the proposed method is shown by a red circle in the corresponding video frame. The largest mouth-to-estimate distance error is 128 pixels (number 3). This corresponds to $\approx 1.7^\circ$ error in azimuth and $\approx 5.3^\circ$ error in elevation. For comparison, results obtained with a face detection algorithm [16] are shown with a yellow square. While this method correctly localized the faces in 10 out of 12 examples, it failed to detect faces #8 and #9 due to partial occlusions. It also featured a number of false detections.

[2]The training data, test data, and video examples are available at http://perception.inrialpes.fr/people/Deleforge/AVASM_Dataset/

[3]A linear dependency was observed in practice.

5. CONCLUSION

We presented a method that maps sound-source directions onto images in order to achieve audio-visual alignment, e.g., speech-to-face association. The direction of a sound is estimated both in azimuth and elevation (2D localization), whereas the vast majority of state-of-the-art techniques estimate only the azimuth (1D localization). 2D localization is absolutely necessary in order to align audio signals with visual features. The method relies on a regression framework [15] that learns the parameters of a transformation from binaural spectrograms to pixel coordinates. While the offline training needs full-spectrum sounds (white noise), the online localization has a closed-form solution and can be used to locate any sound type, including natural sparse-spectrum sounds. Another advantage of our method is that it requires neither explicit representation of the microphone/camera geometry [13, 14], nor long audio-visual sequences [3]. Proper alignment between acoustic spectrograms and images paves the road towards audio-visual data association for advanced multimodal processing and scene understanding.

6. REFERENCES

[1] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *Journal of Acoustical Society of America*, vol. 109, no. 6, pp. 3007–3020, 2001.

[2] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1642–1651, 2011.

[3] D. Segev, Y. Y. Schechner, and M. Elad, "Example-based cross-modal denoising," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 486–493.

[4] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 96 –108, January 2007.

[5] S. M. Naqvi, Y. Miao, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895 –910, oct. 2010.

[6] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1260–1273, 2002.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306 – 1326, sept. 2003.

[8] J. Barker and S. Xu, "Energetic and informational masking effects in an audiovisual speech recognition system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 3, pp. 446–458, March 2009.

[9] A. Deleforge and R. Horaud, "2D sound-source localization on the binaural manifold," in *IEEE International Workshop Machine Learning for Signal Processing.*, Santander, Spain, 2012.

[10] H. Viste and G. Evangelista, "On the use of spatial cues to improve binaural source separation," in *Proc. Int. Conf. on Digital Audio Effects*, 2003, pp. 209–213.

[11] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[12] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1503–1512, 2012.

[13] V. Khalidov, F. Forbes, and R. Horaud, "Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target," in *IEEE Workshop on Multimedia Signal Processing*, 2013.

[14] X. Alameda-Pineda and R. Horaud, "A geometric approach to sound source localization from time-delay estimates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 6, pp. 1082–1095, June 2014.

[15] A. Deleforge, F. Forbes, and R. Horaud, "High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables," *Statistics and Computing*, 2014.

[16] P. Viola and M. J Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," Tech. Rep. NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

[18] P. Aarabi, "Self-localizing dynamic microphone arrays," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 32, no. 4, pp. 474–484, 2002.