# CONTINUOUS VISUAL SPEECH RECOGNITION FOR AUDIO SPEECH ENHANCEMENT

*Eric Benhaim*[⋆†], *Hichem Sahbi* [⋆]

[⋆] Telecom ParisTech
CNRS-LTCI
46 rue Barrault, 75013 Paris, France

*Guillaume Vitte*[†]

[†] Parrot S.A.
174 quai de Jemmapes, 75010 Paris, France

**ABSTRACT**

We introduce in this paper a novel non-blind speech enhancement procedure based on visual speech recognition (VSR). The latter is based on a generative process that analyzes sequences of talking faces and classifies them into visual speech units known as visemes.
We use an effective graphical model able to segment and label a given sequence of talking faces into a sequence of visemes. Our model captures unary potential as well as pairwise interaction; the former models visual appearance of speech units while the latter models their interactions using boundary and visual language model activations. Experiments conducted on a standard challenging dataset, show that when feeding the results of VSR to the speech enhancement procedure, it clearly outperforms baseline blind methods as well as related work.

***Index Terms—*** Visual speech recognition, probabilistic graphical model, belief propagation, model-based speech enhancement

## 1. INTRODUCTION

Over the past twenty years many authors have focused on audio-only methods for speech enhancement. The majority of these methods are based on generalized signal-to-noise ratio (SNR) dependent weighting rules and their accuracies degrade with increasing levels of noise in different environments [1]. Several studies support that speech perception is a multi-modal process which is highly influenced by articulatory movements of speakers' faces. One of the most popular examples that exhibits the multimodal nature of speech perception is known as the McGurk effect [2]: this illusion shows that when a voice saying /ba/ was presented with a face articulating /ga/ most subjects heard /da/. It is therefore admitted that visual speech analysis is essential in order to enhance audio processing systems, especially when the underlying acoustic signals are captured in noisy environments [1]. Recently, many works have focused on HMM-based audiovisual speech methods [3] within a Wiener filtering framework [4]. The growing interest in this research area reflects the need to design robust visual speech analyzer for real-world application, including speech signal enhancement.

In this work, we propose a new framework for continuous visual speech recognition (VSR) based on probabilistic graphical models. Our goal is to effectively discriminate and decode visual speech units in continuous talking sequences for audio speech enhancement. To this end, we propose a viseme-based speech enhancement procedure able to deal with challenging and highly degraded continuous speech.

This work includes two main contributions:

- We propose a unified probabilistic framework that simultaneously recognizes and delimits boundaries of visual units in continuous speech. Our decoding scheme is based on a probabilistic graphical model that (i) explores in an efficient way the search space of possible speech units as well as their boundaries and then (ii) scores and selects the most likely configurations according to speech segment local evidence and contextual constraints.

- We introduce a viseme-based speech enhancement procedure where VSR is used as a front-end of a speech separation algorithm. Our solution is based on a popular compositional model.

The rest of this paper is organized as follows: Section 2 provides the general framework of visual speech recognition using probabilistic graphical model. Section 3 describes the benefit from using visual speech units in order to accurately monitor speech enhancement. Experiments and results obtained are reported in section 4, before concluding in section 5.

## 2. VISUAL SPEECH RECOGNITION

Visemes are visual speech units associated to phonemes in spoken languages. As phonemes are sometimes difficult to distinguish, especially in noisy environments, visemes provide a complementary information that enhances discrimination between speech units. In practice, visemes result from grouping phonemes with similar visual appearances.

Considering $\mathcal{P}$ as a fixed set of 41 phoneme labels, we use a surjective mapping $\pi : \mathcal{P} \rightarrow \mathcal{V}$, with $\pi$, $\mathcal{V}$ being resp. a mapping and a set of visemes taken from [5] (see table 1).

| JEFFERS MAP [5] | | | |
|---|---|---|---|
| *Viseme* | *Phonemes* | *Viseme* | *Phonemes* |
| A | /f/ /v/ | H | /s/ /z/ |
| B | /er/ /ow/ /r/ /w/ /uh/ /uw/ | I | /aa/ /ae/ /ay/ /eh/ /ah/ /ey/ /ih/ /iy/ /y/ /ax/ |
| C | /b/ /p/ /m/ | | |
| D | /aw/ | J | /d/ /l/ /t/ /n/ /dx/ |
| E | /dh/ /th/ | | |
| F | /ch/ /jh/ /sh/ | K | /k/ /g/ /ng/ /hh/ |
| G | /oy/ /ao/ | S | /sil/ |

**Table 1**. Linguistic-based "many-to-one" viseme mapping used in our experiments. Phonemes are clustered into 11 viseme classes plus silence viseme.

Our goal is to tackle continuous speech recognition by finding a sequence of viseme labels and their boundaries $(\mathbf{V}^*, \boldsymbol{\gamma}^*)$, that maximizes *a posteriori* probability

$$(\mathbf{V}^*, \boldsymbol{\gamma}^*) = \text{argmax}_{\mathbf{V},\boldsymbol{\gamma}} P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X}), \tag{1}$$

here $\mathbf{X} = [x^1, x^2, \ldots, x^T]$ is a sequence of successive multidimensional input observations (corresponding to a given talking person) and $\mathbf{V} = [v^1, v^2, \ldots, v^n]$ is the underlying (unknown) sequence of visual speech unit labels with each $v^i \in \mathcal{V}$. We also define $\boldsymbol{\gamma} = [\gamma^1, \gamma^2, \ldots, \gamma^n]$ as $n$ (unknown) positive values that delimit time intervals of each visual speech unit in $\mathbf{V}$ with $\gamma^0 < \gamma^1 < \cdots < \gamma^n = T$ and $\gamma^0 = 0$; so the time interval associated to $v^i$ is defined as $]\gamma^{i-1}, \gamma^i]$.

Graphical models provide a natural way of encoding dependencies and interactions between visual speech segments, as a left-to-right chain structure (see figure 1). A sequence $\mathbf{X}$ of successive input observations is divided into *speech atomic segments* as $\mathbf{X} = \{\mathbf{x}_w^i\}_i$ where $\mathbf{x}_\omega^i = [x^{i\cdot\omega}, \ldots, x^{(i+1)\cdot\omega}[$ is a temporal window of successive observations of length $\omega$. In what follows, $\mathbf{x}_\omega^i$ is referred to as *attribute*. In practice we take $\omega = 2$.

We model a distribution over labeling of attributes via a factor graph where each node is associated to a random variable that describes a given attribute. This model makes it possible to propagate beliefs through nodes based on their local evidences and their interactions (see sections 2.1, 2.2).

$$P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X}) \propto \prod_i \underbrace{\phi_i(\mathbf{x}_\omega^i, v^i)}_{\text{local evidence}} \prod_{i,j} \underbrace{\psi_{ij}(v^i, v^j)}_{\text{pairwise potential}} \tag{2}$$

We use Belief Propagation (BP) to predict the *optimal* configuration of labels $\mathbf{V} = [v^1, v^2, \ldots, v^n]$ associated to $\boldsymbol{\gamma} = [\gamma^1, \gamma^2, \ldots, \gamma^n]$ by minimizing an objective function which trades off unary potential and binary interaction terms [6].

### 2.1. Kernel-based unary potential

Considering $\mathcal{X}$ as the union of all possible sequences taken from the same distribution as $\mathbf{X}$, we define $\mathcal{T}_\omega = \{(\mathbf{x}_\omega^i, v^i)\}_i$ as a training set with each $\mathbf{x}_\omega^i$ corresponds to an *attribute* instance

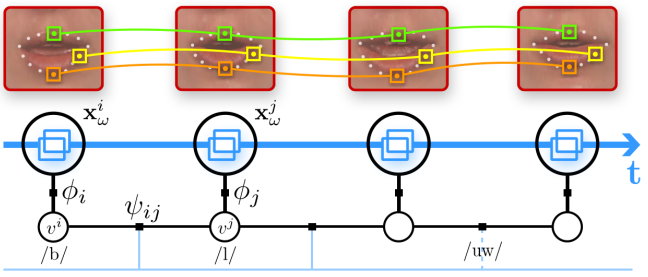



**Fig. 1**. Illustration of the graphical model

i.e. a well delimited subsequence[1] and $v^i$ its viseme label in $\mathcal{V}$ (taken from a well defined ground truth).

Multi-class SVMs use a mapping $\Phi$, that takes data from the input space to a high (possibly infinite) dimensional space and find an optimal separating hyperplane in that high dimensional space. Given classes $\{v \in \mathcal{V}\}$, training is achieved by solving the following quadratic programming problem

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{b},\xi} \frac{1}{2} \sum_{v \in \mathcal{V}} \langle \boldsymbol{w}_v, \boldsymbol{w}_v \rangle + \sum_{i=1}^{|\mathcal{T}_\omega|} \xi^i \\ &\text{s.t} \quad \xi^i = \max_{v \in \mathcal{V} \setminus v^i} l(f_{v^i}(\mathbf{x}_\omega^i) - f_v(\mathbf{x}_\omega^i)), \ \forall\, i, \end{aligned} \tag{3}$$

here $f_v(\mathbf{x}_\omega^i) = \langle \boldsymbol{w}_v, \Phi(\mathbf{x}_\omega^i) \rangle + \boldsymbol{b}_v$ with $\boldsymbol{w}_v$ and $\boldsymbol{b}_v$ being respectively hyperplane normal and bias associated to a given class $v \in \mathcal{V}$ and $\boldsymbol{w} = \{\boldsymbol{w}_v\}_v$, $\boldsymbol{b} = \{\boldsymbol{b}_v\}_v$, $\xi = \{\xi^i\}_i$ and $l(.)$ is a convex loss function. Note that, in practice, we use string kernel maps for $\Phi$ [7]. Details about the design of these kernel maps, out of the main scope of this paper, are deliberately omitted and can be found in [7].

We define attribute unary potential by turning the scores provided by SVMs for different viseme classes into class probability distribution using the method in [8]. The latter is based on the Levenberg-Marquardt algorithm that uses an additional sigmoid in order to define class probability distribution.

$$\phi_i(\mathbf{x}_\omega^i, v) \propto \left(1 + \exp\{A_v f_v(\mathbf{x}_\omega^i) + B_v\}\right)^{-1}, \tag{4}$$

here $A_v$ and $B_v$ are optimized once by minimizing a local negative log-likelihood on a training set.

### 2.2. Pairwise potential

We propose binary interaction terms that take into account two types of contextual constraints: temporal context and semantic context. The purpose of this new designed pairwise potential is (i) to enforce the consistency between temporally close attribute labels and (ii) to enforce agreement between different linked attribute labels (i.e. in potential boundaries) based on high-level viseme language model.

[1] any subsequence of observations, taken from a given sequence in $\mathcal{X}$ but corresponds to a single viseme, is decomposed into $\omega$-length window samples with an overlap factor of 0.5.

**Boundary activation.** Let's define a boundary activation function as

$$\psi_{\Delta_{ij}} = \sigma(s_b - \lambda_{ij}) \tag{5}$$

with $s_b$ being a boundary detection score (described below) and $\sigma(x) = 1/(1+\exp(-1.4\cdot x))$ is a logistic function. In the above equation, parameters $\{\lambda_{ij}\}_{ij}$ are different for each possible label transition $v^i \rightarrow v^j$ in order to perform context *re-scoring*. In practice, $\lambda_{ij}$ is set to the proportion of pairs of viseme labels present in a training dataset.

We compute HOGHOF[9] spatio-temporal descriptors along trajectories of locations $\{p\}_p$ tracked around the mouth. A boundary detection score $s_b$ is obtained by applying an SVM. The latter is trained with histogram intersection kernel on a population of boundary (positive) and non-boundary (negative) features. A feature is defined as $\Delta_{ij} = [\Delta^p_{ij}]_p$, with $\Delta^p_{ij}$ being the $l_2$-norm of the difference between the HOGHOF descriptors taken from location $p$ at two consecutive frames.

**Viseme language model.** In order to build the viseme language model, we automatically generate transcriptions (at the viseme level) from a large corpus of data. For that purpose, we use the Carnegie Mellon pronouncing dictionary[2]. Jeffers [5] mapping is applied in order to convert the phonetic transcriptions into viseme sequence. The bigram viseme language model corresponds to a smoothed probability distribution $P$ on pairs of viseme labels, estimated by parsing and counting the frequencies of all pairs of viseme labels present into the training corpus. Considering $\psi_{lm}(v^i, v^j) = P(v^i, v^j)$, we write the pairwise potential function as

$$\psi_{ij}(v^i, v^j) = \begin{cases} c \cdot \psi_{\Delta_{ij}} + (1-c) \cdot \psi_{lm}(v^i, v^j) & v^i \neq v^j \\ 1 - \psi_{\Delta_{ij}} & v^i = v^j \end{cases} \tag{6}$$

This pairwise potential function is used in order to propagate attribute labels through the structure of the graphical model according to the boundary activation function $\psi_{\Delta_{ij}}$ and the language model function $\psi_{lm}$ while ensuring temporal label consistency; as shown in this model, high values of these two functions, encourage label transition whereas smaller values (of $\psi_{\Delta_{ij}}$ in particular) keep the same label. Note that $c = 0.76$ and this provides good behavior of our model.

We use sum-product message passing to efficiently and effectively solve the inference problem (1) (see [6] for more details). This algorithm iteratively updates labels of random variables associated to the graphical models according to the constraints imposed by the proposed unary and pairwise potential.

## 3. MULTIMODAL SPEECH ENHANCEMENT

### 3.1. Compositional model

Our framework is based on a popular *compositional* model, in which the magnitude spectra of an audio signal is modeled by a linear combination of a set of spectral bases that represent characteristic spectro-temporal patterns.

[2]http://www.repository.voxforge1.org/download/SpeechCorpus

We consider a dictionary of speech atoms $\mathbf{D}_s = [\mathbf{D}_y]_{y\in\mathcal{P}}$ with each $\mathbf{D}_y$ associated to the phoneme $y \in \mathcal{P}$, and a dictionary of noise atoms $\mathbf{D}_n$. Actually, the atoms in the dictionary span three audio-frames. In order to set out dictionary atoms[3] from phonetically labelled training material, we use *examplar-based* characterization [10].

Assuming additivity of speech and noise, each noisy audio observation $x$, taken from a given sequence of successive audio-frames, can be approximated as a linear combination of atoms, which manages to reveal the most likely atomic components as such

$$x \simeq \sum_{y\in\mathcal{P}} \mathbf{D}_y\alpha_y(x) + \mathbf{D}_n\alpha_n(x) = \mathbf{D}\alpha(x) \tag{7}$$

where $\alpha_y(x)$ and $\alpha_n(x)$ resp. define the non-negative *activation weights* for each speech and noise atoms. We note $\mathbf{D} = [\mathbf{D}_s\mathbf{D}_n]$ the combination of speech and noise dictionaries, and $\alpha(x) = [\alpha_1(x)' \ldots \alpha_{|\mathcal{P}|}(x)'\alpha_n(x)']'$ the combination of non-negative activation vector into a single vector.

We apply non-negative matrix factorisation to construct a *sparse representation* of the observed mixture, by minimizing the number of active atoms. Again, $\alpha_y(x)$ and $\alpha_n(x)$ resp. describe the spectral realization of each phoneme class $y$ and the noise class, contributing to the observed mixture.

### 3.2. VSep: Viseme-dependent separation

Given $\mathbf{X} = [x^1, x^2, \ldots, x^T]$ a sequence of successive input observations corresponding to a given speaker, the proposed continuous VSR algorithm finds the optimal decoded sequence of visemes $\mathbf{V} = [v^1, v^2, \ldots, v^n]$ and correponding time interval $\gamma = [\gamma^1, \gamma^2, \ldots, \gamma^n]$ as discussed in section 2. Knowing the viseme $v^i \in \mathcal{V}$ for any segment of speech enables us to restrict the bases required to compose instances of the current pronounced phoneme. Thus for each noisy audio observation $x^i \in ]\gamma^{i-1}, \gamma^i]$, the compositional model can be written as

$$x^i \simeq \sum_{y\in\mathcal{P}/\pi(y)=v^i} \mathbf{D}_y\alpha_y(x^i) + \mathbf{D}_n\alpha_n(x^i) \tag{8}$$

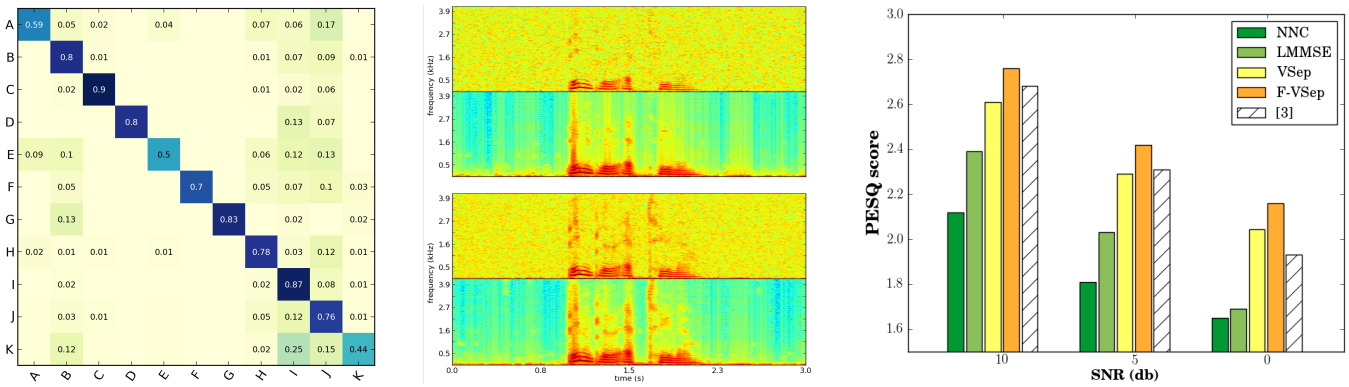
An estimate of the clean speech spectrum is derived from $\alpha(x^i)$ by a Wiener-type filtering as $\mathbf{D}_s\alpha_s(x^i)/\mathbf{D}\alpha(x^i)$ with $\mathbf{D} = [\mathbf{D}_s\mathbf{D}_n]$ and $\alpha_s(x^i) = [\delta[\pi(y) - v^i] \cdot \alpha_y(x^i)']'_{y\in\mathcal{P}}$ where $\delta$ is the Kronecker delta function.
Finally the clean waveform is estimated using the reconstructed speech spectrogram by combining it with the noisy phase, and applying the inverse discrete fourier transform.

## 4. EXPERIMENTS

### 4.1. Experimental condition

We use the GRID audio-visual sentence corpus [11] in order to evaluate our method. Two male and two female speakers

[3]Phoneme specific - and noise - atoms are resp. obtained by drawing random spectral vectors from segments of corresponding phonemes samples, and segments of non-speech corrupted signal.




(a) (b1:top) (b2:bottom) (c)

**Fig. 2**. These figures show experiments on the GRID set. Viseme decoding accuracy is shown in **(a)**. Noisy and VSep-based filtered spectrum of the GRID sentence *"bin blue at f two now"* with added white noise at 0dB SNR and 10dB SNR are resp. shown in **(b1)** and **(b2)**. Speech enhancement PESQ scores comparing blind and non-blind filtering methods are shown in **(c)**.

recordings were divided into a training and a testing sets. It was recorded at 25 fps with a spatial resolution of $720 \times 576$ pixels. The acoustic speech for each utterance is encoded at 256 kbps with a sampling rate of 44.1kHz. The sentences follow a unique grammatical structure each one composed of a combination of six word commands: verb, color, preposition, letter, digit, and coda (e.g. *"bin blue at f two now"*). Phonetic transcriptions were automatically generated using a Viterbi forced alignment procedure [12].

The test sets for all speakers are artificially distorted with white noise at SNR ranging from 10dB to 0dB. As discussed in section 3.2, viseme-based speech separation is achieved using a dictionary of 200 atoms per phoneme class.

To asses the quality of the enhanced speech, we used PESQ [13] as a objective quality measurement which correlates with perceived speech quality.

### 4.2. Results and comparison

Figure 2-a shows the confusion matrix obtained when segmenting and labelling visual speech units. Our proposed graphical model effectively decodes visemes with an accuracy that reaches $79.4\%$, in comparison with $71.1\%$ obtained when the estimated viseme posteriors (as described in section 2.1) are fed into a conventional HMM/GMM system. The authors of [14] provide HMM-based recognition results and show an accuracy of $73.1\%$ for speaker-dependent experiments.
It appears that our proposed architecture indeed capture the temporal dynamics of visual speech of which our graphical model makes full uses. Note that recent methods [15, 16, 17] were introduced for visual-only speech recognition task but none were tested on continuous visual speech. The systems were used for classifying isolated words/phrases, as in [3] where authors reported visual-only GRID-word recognition accuracy of $84.1\%$. It is unknown whether these methods are suitable for recognition at the viseme level.

Figure 2-b shows spectrograms of utterance *"bin blue at f two now"* distored with white noise resp. at an SNR of 0dB (b1) and 10db (b2), and after VSep filtering. We note that VSep filtering has successfully removed large amounts of noise at all SNRs, in non-speech and speech periods.

Figure 2-c presents objective quality measure for different blind (i.e., without knowing viseme classes) and non-blind speech enhancement procedures. It also relates the quality of the unprocessed signal (NNC) as reference.
-The results show that VSep outperformed standard well-established audio-only speech enhancement method, the log minimum mean square error (LMMSE).
-We compare our framework with a recent successful method called *twin-HMM* [3] which used audiovisual GRID-word recognition and a synthesis model. Both results show similar performance[4], although VSep tend to be more robust in very challenging condition (0dB SNR).
-Viseme decoding performance and quality of speech enhancement was correlated running a "controlled" procedure (F-Vsep). We used forced-align viseme transcription when running viseme-based speech enhancement. We note that F-VSep slightly improves speech separation and offers best objective quality measures. It reflects the importance to propose accurate VSR systems, especially in large-vocabulary continuous speech conditions.

## 5. CONCLUSION

We introduced in this paper a unified probabilistic framework that recognizes and simultaneously delimits boundaries of visual units in continuous speech. We proposed a kernel-based unary potential and a pairwise interaction terms that capture visual speech segment local evidence and contextual constraints. A viseme-based speech enhancement procedure is presented and state-of-the-art performances in challenging noisy conditions has been shown.
As a future work we are investigating the design of more complex phoneme-to-viseme relationship in order to handle the natural asynchrony of audio-visual speech.

[4]Since training/testing sets from [3] are unknown, only behaviour can be compared.

## 6. REFERENCES


[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Audio-Visual Speech Processing*, 2004.

[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 1976.

[3] Ahmed Hussen Abdelaziz, Steffen Zeiler, and Dorothea Kolossa, "Twin-hmm-based audio-visual speech enhancement," in *ICASSP*. IEEE, 2013.

[4] Ibrahim Almajai and Ben Milner, "Visually derived wiener filters for speech enhancement," *Audio, Speech, and Language Processing*, 2011.

[5] Janet Jeffers and Margaret Barley, *Speechreading (lipreading)*, Thomas Springfield, 1971.

[6] Jonathan S Yedidia, William T Freeman, and Yair Weiss, "Understanding belief propagation and its generalizations," *Artificial intelligence*, 2003.

[7] Eric Benhaim, Hichem Sahbi, and Guillaume Vitte, "Designing relevant features for visual speech recognition," in *ICASSP*. IEEE, 2013.

[8] John Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, 1999.

[9] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, Cordelia Schmid, et al., "Evaluation of local spatio-temporal features for action recognition," in *BMVC*, 2009.

[10] Jort F Gemmeke and Tuomas Virtanen, "Noise robust exemplar-based connected digit recognition," in *ICASSP*. IEEE, 2010.

[11] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, 2006.

[12] Ahsanul Kabir, Mircea Giurgiu, and Jon Barker, "Robust automatic transcription of english speech corpora," in *Communications (COMM)*. IEEE, 2010.

[13] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*. IEEE, 2001.

[14] Xu Shao and Jon Barker, "Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment," *Speech Communication*, 2008.

[15] Yuru Pei, Tae-Kyun Kim, and Hongbin Zha, "Unsupervised random forest manifold alignment for lipreading," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 129–136.

[16] Eng-Jon Ong and Richard Bowden, "Learning temporal signatures for lip reading," in *Computer Vision Workshops ICCV*. IEEE, 2011.

[17] Z. Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *CVPR*. IEEE, 2011.