CONTINUOUS VISUAL SPEECH RECOGNITION FOR MULTIMODAL FUSION

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

It is admitted that human speech perception is a multimodal process that combines both visual and acoustic informations. In automatic speech perception, visual analysis is also crucial as it provides a complementary information in order to enhance the performances of audio systems especially in highly noisy environments.
In this paper, we propose a unified probabilistic framework for speech unit recognition that combines both visual and audio informations. The method is based on the optimization of a criterion that achieves continuous speech unit segmentation and decoding using a learned (joint) phonetic-visemic model. Experiments conducted on the standard LIPS2008 dataset, show a clear and a consistent gain of our multimodal approach compared to others.

Index Terms— Visual speech unit recognition, multi-class support vector machines, multimodal segmentation.

1. INTRODUCTION

Several studies support that speech perception is a multimodal 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 [1]: 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 automatic speech recognition (ASR) systems, especially when the underlying acoustic signals are captured in noisy environments [2].

Recently, many works have focused on visual speech recognition (VSR) also known as lipreading. The growing interest in this research area reflects the need to design robust visual speech analyzers for real-world applications, including human machine interaction for multimodal remote control, assisting experts in decoding video evidences, monitoring public places with video surveillance, or speech signal enhancement dedicated to in-car communication.

Continuous visual speech recognition is a temporal decoding of sequences of visual speech units known as visemes[1]. While the numerous existing ASR solutions range from speaker dependent isolated word recognition, to speaker independent phoneme recognition, there is still no well-defined baseline systems for continuous VSR; existing recognition systems are restricted to digits [3, 4, 5], letters [6, 7], words [6], or short phrases [8, 9, 10]. Only a few work presented continuous VSR performance on short vocabulary sentences [11, 12]. Within this context, authors have mainly focused in the past decade, on designing relevant visual features that better capture speech induced variability rather than the appearance of the speaker [13], while being speaker independent [3, 5, 10]. This is still considered as an open problem due to large speaker inherent variability in lip-motion and appearance.
Despite the increasing interest in this domain, the challenge of continuous VSR remains threefold. Firstly, it is still unclear what definition of visual speech units (i.e., visemes) should be used for real-world applications and in practice several phoneme-to-viseme relationships have been proposed (see for instance [12, 14, 15, 16]) with some advantages and insufficiencies [17]. Secondly, building lipreading systems requires annotated audio-visual continuous speech datasets which are scarce and the few existing ones require tedious and error prone manual generation of the ground truth. Furthermore suitable datasets are expected to have diversity in speakers and the vocabulary used in uttered sentences. Among the few existing databases, neither AV-TIMIT [12] nor AV-ViaVoice [15] are publicly available, and XM2VTS [18] is not free. Fortunately, the large vocabulary LIPS2008 database [19], originally designed for speech synthesis purposes, is available and constitutes a suitable alternative. Finally, considering the requirement of the targeted communication framework, lipreading systems should involve effective classifiers able to encode time-varying speech utterances and efficient decoding schemes for speech segmentation.

In this paper, we propose a novel learning framework for continuous VSR based on support vector machines (SVMs)[2]. Our method is multimodal and unifies the problem of visual and acoustic speech unit recognition using a probabilistic framework. We will show that our visual model is able to reduce phoneme class confusion due to acquisition conditions as well as signal variability. In order to tackle these issues, our work includes the following contributions
-We propose a unified probabilistic framework that *simultaneously* recognizes and delimits boundaries of visual and acoustic units in continuous speech. Our decoding scheme is based on a 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 configuration.
-We design a scoring function based on a Bayesian classifier that combines the output of SVMs with an a priori language model that captures joint statistics of visemes and phonemes. For that purpose, we extend this study by comparing different viseme definitions.

The rest of this paper is organized as follows: Section 2 provides speech unit definitions and gives a general formulation of the recognition task. Section 3 describes our visual learning framework. Section 4 establishes the multimodal fusion scheme and presents an efficient sequence decoding procedure. Experiments and results obtained are reported in section 5, before concluding in section 6.

[1] Visemes are visual units of speech associated to phonemes in spoken languages.

[2] The choice of SVMs was also motivated by their good generalization capability, compared to other models, in order to handle few training examples in high dimensional spaces.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

JEFFERS MAP [14]	
Viseme	*Phonemes*
A	/f/ /v/
B	/er/ /ow/ /r/ /w/ /uh/ /uw/
C	/b/ /p/ /m/
D	/aw/
E	/dh/ /th/
F	/ch/ /jh/ /sh/
G	/oy/ /ao/
H	/s/ /z/
I	/aa/ /ae/ /ay/ /eh/ /ah/ /ey/ /ih/ /iy/ /y/ /ax/
J	/d/ /l/ /t/ /n/ /dx/
K	/k/ /g/ /ng/ /hh/
S	/sil/

NETI MAP [15]	
Viseme	*Phonemes*
V1	/ao/ /ah/ /aa/ /oy/ /aw/ /er/ /hh/
V2	/uw/ /uh/ /ow/ /oh/
V3	/ae/ /eh/ /ey/ /ay/
V4	/ih/ /iy/ /ax/
A	/l/ /r/ /y/
B	/s/ /z/
C	/t/ /d/ /n/
D	/sh/ /ch/ /jh/
E	/p/ /b/ /m/
F	/th/ /dh/
G	/f/ /v/
H	k/ /g/ /w/ /ng/
S	/sil/

MPEG-4 MAP [16]	
Viseme	*Phonemes*
V1	/b/ /p/ /m/
V2	/f/ /v/
V3	/th/ /dh/
V4	/d/ /dx/ /t/
V5	/k/ /g/ /ng/ /hh/
V6	/sh/ /jh/ /ch/
V7	/s/ /z/
V8	/n/ /l/
V9	/r/ /er/
V10	/y/ /aa/ /ao/ /aw/ /oy/ /ah/ /ax/
V11	/eh/ /ey/ /ae/ /ay/
V12	/iy/ /ih/
V13	/ow/
V14	/uh/ /uw/ /w/
S	/sil/

HAZEN MAP [12]	
Viseme	*Phonemes*
OV	/ax/ /ih/ /iy/ /dx/
BV	/ah/ /aa/
FV	/ae/ /eh/ /ay/ /ey/ /hh/
RV	/aw/ /uh/ /uw/ /ow/ /ao/ /oy/ /w/
L	/l/
R	/r/ /er/
Y	/y/
LB	/b/ /p/
LCl	/m/
AlCl	/s/ /z/ /n/
Pal	/jh/ /ch/ /sh/
SB	/t/ /d/ /th/ /dh/ /k/ /g/
LFr	/f/ /v/
VlCl	/ng/
S	/sil/

Table 1. The tables show four "many-to-one" viseme mappings tested in our experiments. These mappings are based on linguistic and/or data-driven methods, and differ in their number of viseme classes: from 11 to 14, plus silence viseme S. Note that we re-defined all these mappings on the same input set $\mathcal{P}$ of 41 phonemes.

2. PROBLEM FORMULATION

In this section, we introduce different speech units[3] and the underlying mapping functions. We also introduce our problem formulation that allows us to tackle continuous speech unit recognition.

2.1. Speech Unit Mapping

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. This grouping is usually defined from human experts' knowledge (and hence varies from one expert to another [14, 16]) or can be inferred by learning from data [12, 15].
Several *many-to-one* mappings exist in the literature without universal agreement on the exact number of visemes needed to accurately describe visual speech information. Recently more complex *many-to-many* relationships between visemes and phonemes have been defined and applied to computer-based facial animation [17]. However for applications such as speech enhancement and speech unit recognition, straightforward many-to-one mappings, between visual and acoustic units, are preferred.

Considering $\mathcal{P}$ as a fixed set of 41 phoneme labels, we use a surjective mapping $\psi : \mathcal{P} \rightarrow \mathcal{V}$, with ψ, $\mathcal{V}$ being resp. a mapping and a set of visemes taken from one of the following: jeffers [14], neti [15], mpeg-4 [16], and hazen [12]. Table 1 presents these four "many-to-one" viseme mappings which are used in our experiments. Note that we translate all these mappings with a reduced set $\mathcal{P}$ of 39 symbols corresponding to English phonemes [20] commonly used in phoneme recognition and we added phonemes /ax/ and /ao/ to this set. These phonemes belong originally to classes /ah/ and /aa/ respectively but they appear in distinct viseme classes for some mappings.

2.2. Continuous Speech Unit Recognition

Again, the main purpose of VSR is to support and enhance acoustic-based systems in challenging environments. Resulting from the non unicity of the phoneme-to-viseme mappings (see again Table 1), viseme classes are different and it is meaningless to evaluate and compare performance of speech recognition using viseme classes as a target. However, as all these mappings are defined on the same input (phoneme) set, it is preferable to evaluate their performance by considering phonemes as our targeted classes. Thereby, as will be shown through this paper, we propose a multimodal speech unit decoding algorithm that unifies both phoneme and viseme based models.

Our goal is to tackle continuous speech recognition by finding a sequence of speech unit labels and their boundaries $(\mathbf{Y}^*, \boldsymbol{\gamma}^*)$, that maximizes *a posteriori* probability $P(\mathbf{Y}^*, \boldsymbol{\gamma}^*|\mathbf{X})$, here $\mathbf{X} = [x^1, x^2, \ldots, x^T]$ is a sequence of successive input observations (corresponding to a given talking person) and $\mathbf{Y} = [y^1, y^2, \ldots, y^n]$ is the underlying (unknown) sequence of speech unit labels with each $y^i \in \mathcal{P}$. We also define $\boldsymbol{\gamma} = [\gamma^1, \gamma^2, \ldots, \gamma^n]$ as n (unknown) positive values that delimit time intervals of each speech unit in $\mathbf{Y}$ with $\gamma^0 < \gamma^1 < \cdots < \gamma^n = T$ and $\gamma^0 = 0$; so the time interval associated to y^i is defined as $]\gamma^{i-1}, \gamma^i]$.
Considering $\mathbf{V} = [v^1, v^2, \ldots, v^n]$ as a sequence of visual units associated to $\mathbf{Y} = [y^1, y^2, \ldots, y^n]$, with each $v^i \in \mathcal{V}$, we rewrite the posterior probability defined earlier as

$$P(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{X}) \quad = \quad \sum_{\mathbf{V} \in \mathcal{V}^n} P(\mathbf{Y}|\mathbf{V}, \boldsymbol{\gamma}, \mathbf{X}) P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X}), \qquad (1)$$

here $P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X}) \propto P(\mathbf{X}|\mathbf{V}, \boldsymbol{\gamma}) P(\mathbf{V})$, with $P(\mathbf{X}|\mathbf{V}, \boldsymbol{\gamma})$ being the likelihood of a sequence $\mathbf{X}$ given viseme labels in $\mathbf{V}$. The *prior* $P(\mathbf{V})$ corresponds to the probability of a given sequence of viseme labels while $P(\mathbf{Y}|\mathbf{V}, \boldsymbol{\gamma}, \mathbf{X})$ is a joint phoneme/viseme a priori model, whose design is shown later in this paper.

3. VISUAL LEARNING FRAMEWORK

This section describes our visual learning model which consists in a multi-class SVM and a visemic language model that learns speech unit transitions using a large corpus of phonetic transcriptions and their associated visemic maps.

3.1. Discriminative Training with Multi-class SVMs

Considering $\mathcal{X}$ as the union of all possible sequences taken from the same distribution as $\mathbf{X}$ (see Section 2.2), we define $\mathcal{T} = \{(\mathbf{x}_i, v^i)\}_i$ as a training set with each $\mathbf{x}_i$ corresponds to an instance of a well delimited subsequence[4] and v^i its viseme label in $\mathcal{V}$ (taken from a well defined ground truth). Multi-class SVMs use a mapping Φ, that takes data from the input space to a high (possibly infinite) dimensional space and find an optimal separating hyperplane in that high

[3] Again, a speech unit refers to a viseme for video and a phoneme for audio.

[4] i.e., any subsequence of observations taken from a given sequence in $\mathcal{X}$ but corresponds to a single viseme.

dimensional space. Given classes $\{v \in \mathcal{V}\}$, training is achieved by solving the following quadratic programming problem

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

here $f_v(\mathbf{x}) = \langle \boldsymbol{w}_v, \Phi(\mathbf{x})\rangle + \boldsymbol{b}_v$ 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 Φ [5], which are able to transform sequences of varying lengths in $\{\mathbf{x}_i\}_i$ into high dimensional feature vectors. Details about the design of these kernel maps, out of the main scope of this paper, are deliberately omitted and can be found in [5].

Now we turn the scores provided by SVMs for different viseme classes into class probability distribution using the method in [21]. The latter is based on the Levenberg-Marquardt algorithm that uses an additional sigmoid in order to define class probability distribution as $p(v|\mathbf{x}) \propto \left(1 + \exp\{A_v f_v(\mathbf{x}) + B_v\}\right)^{-1}$, here A_v and B_v are optimized once by minimizing a local negative log-likelihood on a training set.

Given a sequence $\mathbf{X}$ of T observations partitioned using $\boldsymbol{\gamma} = [\gamma^1, \ldots, \gamma^n]$ into n subsequences $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, we estimate the posterior probability of any sequence of n viseme labels $\mathbf{V} = [v^1, \ldots, v^n]$ given $\mathbf{X}$ as $P(\mathbf{V}, \gamma|\mathbf{X}) = \prod_{i=1}^{n} p(v^i|\mathbf{x}_i)$, with $\mathbf{x}_i$ being the i^{th} subsequence of $\mathbf{X}$ delimited by $]\gamma^{i-1}, \gamma^i]$.

3.2. Viseme Language Modeling

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[5] which contains more than 130k words. We applied different mappings defined in Table 1, in order to convert the phonetic transcriptions into viseme sequences.

The viseme (l-gram) language model provides the probability $P(\mathbf{V})$ that a given sentence $\mathbf{V} = [v^1, \ldots, v^n]$ is observed as

$$P(\mathbf{V}) = P(v^1)\prod_{k=2}^{n} P(v^k|v^{k-1}, \ldots, v^{k-l+1}). \tag{3}$$

In the above probability, $P(v^k|v^{k-1}, \ldots, v^{k-l+1})$ is estimated by parsing and computing the frequencies of all sequences of l viseme labels present into the training corpus. Notice that this Maximum Likelihood based estimator overestimates the probabilities of l-viseme sequences appearing in the training corpus, while it underestimates those which are not present. Therefore, we apply smoothing [22] in order to re-balance the estimated probabilities.

4. SEGMENTATION AND MULTIMODAL FUSION

In this section, we introduce our main contributions, which allows us (i) to unify viseme and phoneme decoding in a global probabilistic framework and (ii) to approach the segmentation problem for continuous speech.

[5] http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Lexicon/

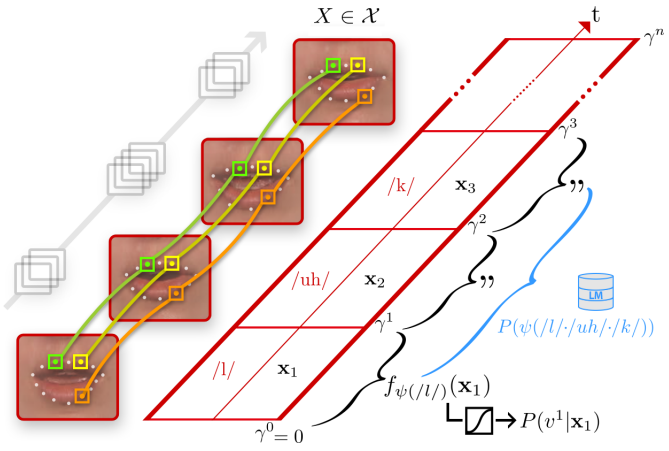

Fig. 1. **Visual learning framework:** Let $\mathbf{X}$ be an input sequence of audio-visual observations. Time intervals $\boldsymbol{\gamma} = [\gamma^1, \ldots, \gamma^n]$ of speech units and corresponding sequence of phoneme labels $\mathbf{Y} = [y^1, \ldots, y^n]$ are provided. $\mathbf{x_i}$ is the i^{th} subsequence of $\mathbf{X}$ delimited by $]\gamma^{i-1}, \gamma^i]$ and $v^i = \psi(y^i)$ its viseme label provided by a mapping function ψ. SVMs are trained for different viseme classes $\{v \in \mathcal{V}\}$ and SVM scores are turned into probabilities (see Section 3.1). A statistical (l-gram) language model is also estimated from phonetic transcriptions (see Section 3.2).

4.1. Phoneme Scoring

We use signal separation for phoneme scoring where each phoneme has a distinct spectral structure. We consider a dictionary of atoms $\mathbf{D} = [\mathbf{D}_y]_y$ with each $\mathbf{D}_y$ associated to the phoneme class $y \in \mathcal{P}$. Using this dictionary, each audio observation x, taken from a given sequence of successive audio-frames, can be approximated as $x \simeq \mathbf{D}\,\boldsymbol{\alpha}(x) = \sum_{y\in\mathcal{P}} \mathbf{D}_y\,\boldsymbol{\alpha}_y(x)$, with $\boldsymbol{\alpha}(x) = [\boldsymbol{\alpha}_1(x)' \ldots \boldsymbol{\alpha}_{|\mathcal{P}|}(x)']'$ being the non-negative activation vector which is normalized such as $\boldsymbol{\alpha}(x)'\boldsymbol{\alpha}(x) = 1$. In this decomposition, each $\boldsymbol{\alpha}_y(x)$ describes the spectral realization of the phoneme class y. In practice, the atoms in the dictionary span three audio-frames. We use *exemplar-based* characterization [23] in order to set our dictionary atoms and apply non-negative matrix factorization to find the coefficients $\boldsymbol{\alpha}(x)$ that describe each observation x as a linear combination of atoms.

Given a sequence $\mathbf{X}$ composed of T audio observations partitioned using $\boldsymbol{\gamma} = [\gamma^1, \ldots, \gamma^n]$ into n subsequences $[\mathbf{x}_1, \ldots, \mathbf{x}_n]$, we estimate the posterior probability of any sequence of n phoneme labels $\mathbf{Y} = [y^1, \ldots, y^n]$ given $\mathbf{X}$ as $\tilde{P}(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{X}) = \prod_{i=1}^{n} \left\|\boldsymbol{\alpha}_{y^i}(\mathbf{x}_i)\right\|_1 / \left\|\boldsymbol{\alpha}(\mathbf{x}_i)\right\|_1$, with $\mathbf{x}_i$ being the i^{th} subsequence of $\mathbf{X}$ delimited by $]\gamma^{i-1}, \gamma^i]$, and $\boldsymbol{\alpha}_{y^i}(\mathbf{x}_i)$ is the activation vector associated to $\mathbf{x}_i$ and to the phoneme class y^i.

4.2. Multimodal Rescoring and Segmentation

We introduce a unified probabilistic framework that combines both the viseme and the phoneme models described earlier in order to rescore speech units and to handle segmentation for a given sequence $\mathbf{X}$. Considering Eq. (1), we rewrite

$$P(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{X}) \quad = \quad \sum_{\mathbf{V}\in\mathcal{V}^n} P(\mathbf{Y}|\mathbf{V})P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X}). \tag{4}$$

The term $P(\mathbf{V}, \boldsymbol{\gamma}|\mathbf{X})$ is estimated as discussed earlier in Section 3, while $P(\mathbf{Y}|\mathbf{V}) = P(\mathbf{Y}, \mathbf{V})/P(\mathbf{V})$ is a *joint* viseme-phoneme language model with $P(\mathbf{Y}, \mathbf{V}) = 1_{\{\psi(\mathbf{Y})=\mathbf{V}\}} \times P(\mathbf{Y})$ (as $\mathbf{V}$ is a deterministic function of $\mathbf{Y}$ defined by the mapping ψ). Similarly to $P(\mathbf{V})$ (see Section 3.2), $P(\mathbf{Y})$ is also estimated.
Now combining the phoneme scoring (in Section 4.1) and the scoring defined by Eq 4, we obtain our unified criterion for segmentation and speech unit rescoring; the best sequence of speech unit labels and its associated segmentation $(\mathbf{Y}^*, \boldsymbol{\gamma}^*)$ correspond to

$$\arg\max_{\mathbf{Y},\boldsymbol{\gamma}} \quad (1-\lambda)\, \tilde{P}(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{X}) + \lambda\, P(\mathbf{Y}, \boldsymbol{\gamma}|\mathbf{X}), \tag{5}$$

here $\lambda \in [0, 1]$. This criterion mixes two terms; the left-hand side term measures the posterior probability of phoneme labels using only the audio information while the second term rescores phoneme labels by applying the visual model as well as the joint viseme phoneme language model.

Optimization. in order to solve (5), we use an efficient greedy algorithm that jointly produces segmentation and speech unit decoding. This algorithm proceeds iteratively by incrementally generating multiple configurations of subsequence boundaries and labelings of a given sequence $\mathbf{X}$. At a given iteration p, the algorithm considers that the best configuration of $[(y^1, \gamma^1) \dots (y^{p-1}, \gamma^{p-1})]$ is known (fixed) and only (y^p, γ^p) is allowed to vary (i.e., $y^p \in \mathcal{P}$ and $\gamma^p \in \{\gamma^{p-1} + l_{min}, \dots, \gamma^{p-1} + l_{max}\}$ with $l_{min} = 2, l_{max} = 16$ in practice); so the best configuration of (y^p, γ^p) is chosen to optimize Eq. 5. The algorithm terminates when all the sequence $\mathbf{X}$ is split into n^* labeled subsequences $(\mathbf{Y}^*, \boldsymbol{\gamma}^*)$, with $n^* \leftarrow p$.

5. EXPERIMENTS

5.1. Evaluation Sets and Settings

We use the LIPS2008 Visual Speech Synthesis Challenge database [19] which contains 278 phonetically balanced sentences spoken by a single, female speaker, in a neutral speaking style. It was recorded at 50 fps with a spatial resolution of 576×720 pixels. The acoustic speech for each utterance is encoded at 16bits/sample with a sampling rate of 44.1kHz. Even and odd sentences are respectively used for training and testing.
We used a combination of string kernels as visual feature mapping Φ in order to measure the similarity as well as the dynamics of visual feature sequences (see [5] for details about kernel design).
We evaluate the priors $P(\mathbf{V})$ as well as the joint viseme-phoneme language model $P(\mathbf{Y}|\mathbf{V})$ using 3-gram language model approximation. Table 2 shows perplexity measures of these language models built from different vocabularies and applied to the test data.

Vocabulary	$\mathcal{P}$	$\mathcal{V}_{\text{jeffers}}$	$\mathcal{V}_{\text{neti}}$	$\mathcal{V}_{\text{mpeg-4}}$	$\mathcal{V}_{\text{hazen}}$
Perplexity	44.1	9.7	11.2	15.9	13.1

Table 2. Perplexity scores.

As discussed in Section 4.1, phoneme scoring is achieved using a dictionary of hundred atoms per phoneme class. Multimodal rescoring and segmentation experiments (see Eq. 5) are conducted with various value of $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$. For a sequence of t outputs, we measure the accuracy of our segmentation and labeling algorithm using $(t - d - s - i)/t$, with t being the number of labels in the ground-truth transcription, and d, s, i being respectively the number of deletions, substitutions and insertions.

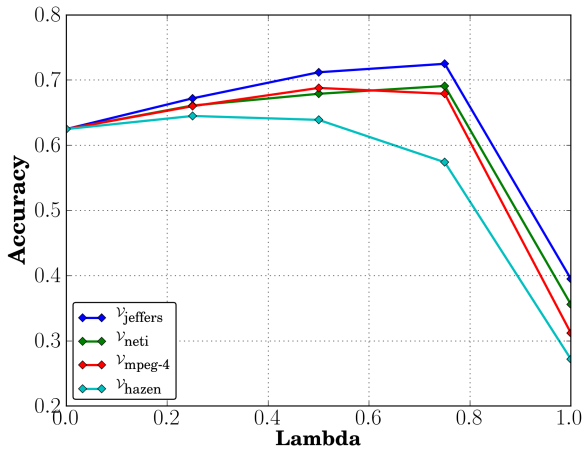

Fig. 2. This figure shows experiments on the LIPS2008 database. Continuous speech unit recognition accuracies w.r.t. different value of λ are shown (see Eq. 5). Four phoneme-to-viseme mappings are compared (see Table 1).

5.2. Results and Comparison

Fig 2 shows the overall speech sequence segmentation performances with respect to different phoneme-to-viseme mappings defined in Table 1. In these results, the baseline corresponds to the accuracy with $\lambda = 0$ (i.e., using only the audio information). These performances are also shown for $\lambda = 1$, which corresponds to the application of the visual model only. In this case, phoneme labels are derived from joint viseme-phoneme statistics applied to the decoded viseme sequence. This explains the drop in performance.
According to these results, the best performances are achieved using Jeffers map [14] with $\lambda = 0.75$ where phoneme class confusion is reduced by more than 10 points and this coincides with the lowest perplexity score (in Table 2). Note that Mpeg-4 [16] and Neti [15] mappings have similar global behaviors as Jeffers map [14]. All these plots show the influence of λ and the number of viseme classes within each mapping; we observe, in particular, that a small number of viseme classes will obviously result into high visual model performance, but a tradeoff is necessary in order to better reduce phoneme class confusion.

6. CONCLUSION

We introduced in this paper a unified probabilistic framework that simultaneously recognizes and delimits boundaries of visual and acoustic units in continuous speech. We proposed a scoring function based on a Bayesian classifier that combines the output of SVMs with an a priori language model that captures joint statistics of visemes and phonemes.
Experiments show that the proposed model is effective and able to reduce substantially phoneme class confusion. Four "many-to-one" phoneme-to-viseme mappings have been compared and the Jeffers mapping provides the best results. As a future work, we are currently investigating the application of our method to acoustic speech enhancement in challenging conditions including noisy car environments. We are also investigating the design of more complex phoneme-to-viseme relationships for VSR application in order to handle the natural asynchrony of audio-visual speech.

7. REFERENCES

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, pp. 746–748, 1976.

[2] 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*, pp. 1306–1326, 2003.

[3] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 3, pp. 423–435, 2009.

[4] M. Gurban and J.P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *Signal Processing, IEEE Transactions on*, vol. 57, no. 12, pp. 4765–4776, 2009.

[5] Eric Benhaim, Hichem Sahbi, and Guillaume Vitte, "Designing relevant features for visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2420–2424.

[6] 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*, vol. 120, pp. 2421, 2006.

[7] S. Cox, R. Harvey, Y. Lan, J. Newman, and B. Theobald, "The challenge of multispeaker lip-reading," in *International Conference on Auditory-Visual Speech Processing*, 2008, pp. 179–184.

[8] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *Multimedia, IEEE Transactions on*, vol. 11, no. 7, pp. 1254–1265, 2009.

[9] Z. Zhou, G. Zhao, and M. Pietikäinen, "Lipreading: a graph embedding approach," in *2010 International Conference on Pattern Recognition*. IEEE, 2010, pp. 523–526.

[10] Eng-Jon Ong and Richard Bowden, "Learning temporal signatures for lip reading," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 958–965.

[11] Luca Cappelletta and Naomi Harte, "Viseme definitions comparison for visual-only speech recognition," in *Proceedings of the European Signal Processing Conference*, 2011, vol. 3.

[12] Timothy J Hazen, Kate Saenko, Chia-Hao La, and James R Glass, "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 235–242.

[13] I. Matthews, T.F. Cootes, J.A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 2, pp. 198–213, 2002.

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

[15] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, Azad Mashari, and Jie Zhou, "Audio-visual speech recognition," in *Final Workshop 2000 Report*, 2000, vol. 764.

[16] Igor S Pandzic and Robert Forchheimer, *MPEG-4 facial animation: the standard, implementation and applications*, Wiley. com, 2003.

[17] Sarah L Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews, "Dynamic units of visual speech," in *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*. Eurographics Association, 2012, pp. 275–284.

[18] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, and Gilbert Maitre, "Xm2vtsdb: The extended m2vts database," in *Second international conference on audio and video-based biometric person authentication*. Citeseer, 1999, vol. 964, pp. 965–966.

[19] Barry-John Theobald, Sascha Fagel, Gérard Gerard, et al., "Lips2008: visual speech synthesis challenge," in *Proceedings of interspeech*, 2008, pp. 2310–2313.

[20] Xuedong Huang, Alejandro Acero, Hsiao-Wuen Hon, et al., *Spoken language processing*, vol. 15, Prentice Hall PTR New Jersey, 2001.

[21] John Platt et al., "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[22] Stanley F Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech & Language*, vol. 13, no. 4, pp. 359–393, 1999.

[23] Jort F Gemmeke and Tuomas Virtanen, "Noise robust exemplar-based connected digit recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4546–4549.