AUDIOVISUAL-TO-ARTICULATORY SPEECH INVERSION USING ACTIVE APPEARANCE MODELS FOR THE FACE AND HIDDEN MARKOV MODELS FOR THE DYNAMICS

Athanassios Katsamanis, George Papandreou and Petros Maragos

School of E.C.E., National Technical University of Athens, Athens 15733, Greece
Email: {nkatsam, gpapan, maragos}@cs.ntua.gr

ABSTRACT

We are interested in recovering aspects of vocal tract's geometry and dynamics from auditory and visual speech cues. We approach the problem in a statistical framework based on Hidden Markov Models and demonstrate effective estimation of the trajectories followed by certain points of interest in the speech production system. Alternative fusion schemes are investigated to account for asynchrony between the modalities and allow independent modeling of the dynamics of the involved streams. Visual cues are extracted from the speaker's face by means of Active Appearance Modeling. We report experiments on the QSMT database which contains audio, video, and electromagnetic articulography data recorded in parallel. The results show that exploiting both audio and visual modalities in a multistream HMM based scheme clearly improves performance relative to either audio or visual-only estimation.

Index Terms— speech inversion, Hidden Markov Models, audiovisual, articulatory, fusion

1. INTRODUCTION

We address the problem of audiovisual-to-articulatory speech inversion, namely the recovery of aspects of the vocal tract shape and dynamics given the speech signal and exploiting visual information from the speaker's face. Speech inversion could potentially allow representing a speech signal by the corresponding vocal tract configuration. This would be quite interesting from a theoretical point of view but also in speech processing applications such as language learning, speech coding and speech therapy.

In this direction, inorporation of the visual modality is considered to be beneficial since there has been a number of studies showing that the speaker's face and the motion of important vocal tract articulators such as the tongue are significantly correlated [1–4]. Motivated by such findings we investigate a statistical framework to recover vocal tract related information by exploiting both the speech signal and visual cues from the speaker's face concurrently recorded.

In [4], the authors explore simple global linear mappings to unveil associations between the behavior of facial data and articulatory data during speech. They conclude that a high percentage (80%) of the variance observed in the vocal tract data can be recovered from the facial data. This conclusion is also verified in [3] on similar data and again by means of global multivariate linear regression. More recently, in [1, 2] articulatory parameters are recovered from facial and audio data either via relevance vector machines or a global linear mapping. These previous studies have shown that, although a global linear mapping is a rough approximation of the underlying complex non-linear interaction between audio-visual features and articulatory positions, it can serve as a first approximation, and also as a baseline system which more advanced techniques have to improve.

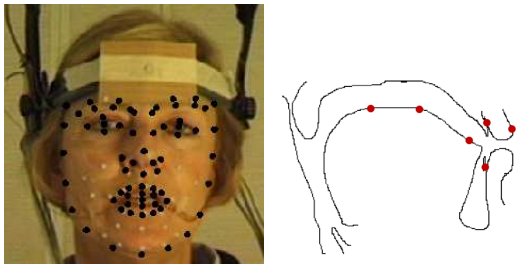

Fig. 1. Qualisys-Movetrack Database. *Left*: Landmarks on the speaker's face have been localized by Active Appearance Modeling and are shown as black dots. White dots are markers glued on the face and tracked during data acquisition. *Right*: Red dots correspond to coils on the speaker's tongue (dorsum, blade and tip from left to right), teeth and lips that have been tracked by electromagnetic articulography. The database also contains speech which is recorded concurrently.

On the other hand, to recover articulatory motion from acoustics only, various sophisticated approaches have been followed. In [5] it is found that Mixture Density Networks perform better than Multilinear Perceptrons in acoustic-to-articulatory inversion. To estimate articulatory trajectories from Mel Frequency Cepstrum Coefficients (MFCCs) derived from the audio signal, a Hidden Markov Model(HMM)-Based Speech Production Model is proposed in [6]. This model allows the imposition of more elaborate constraints to the dynamic behavior of the articulatory parameters that are estimated for given speech acoustics. The HMM scheme is shown to outperform inversion approaches based on codebooks.

In this context, in [7] we proposed the introduction of multistream HMMs to also exploit information from the speaker's face in the inversion task. As far as this visual modality is concerned, in our experiments we utilized the 3D coordinates of the markers glued on the speaker's face and are shown in white in Fig. 1. These were tracked by the Qualisys acquisition system and they are available in the Qualisys-Movetrack (QSMT) dataset [8]. In more realistic scenarios however, only one camera is expected to be recording the motion of the face, which is also expected to be free of any markers. Thus in the present work we propose visual information representation by means of Active Appearance Models. These are generative

This work was supported partially by grant ΠΕΝΕΔ-2003-ΕΔ866 [cofinanced by E.U.-European Social Fund (75%) and the Greek Ministry of Development-GSRT (25%)], by GSRT project DIANOEMA under Grant Image-Sound-Language-35 and partially by European Community FP6 NoE MUSCLE and FP6 FET ASPI (contract no. 021324). We would also like to thank the Speech Communication and Technology Group at KTH for providing us the QSMT database.

models which facilitate effective and robust face modeling. In our case, they allow efficient extraction of visual features which may be used for inversion. To account for asynchrony between the observed modalities we also investigate alternative modeling schemes. Visual and audio dynamics modeling is based on visemes and phonemes correspondingly. Late fusion is then applied to provide articulatory estimates given the combined information. Experiments are reported in the QSMT database.

2. PROPOSED METHOD

2.1. Linear Models for Speech inversion

From a probabilistic point of view, the solution to audiovisual (AV) speech inversion may be seen as the articulatory configuration that maximizes the posterior probability of the articulatory characteristics given the available AV information:

$$p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y}) \quad (1)$$

It would be intuitive to first consider the static case in which both the articulatory and the audiovisual characteristics do not vary with time. The parameter vector $\mathbf{x}$ (n elements) provides a proper representation of the vocal tract. This representation could be either direct, including space coordinates of real articulators, or indirect, describing a suitable articulatory model for example. The Audiovisual parameter vector $\mathbf{y}$ (m elements) should ideally contain all the vocal-tract related information that can be extracted from the acoustic signal on the one hand and speaker's face on the other. Formant values, linear spectral pairs or MFCCs have been applied as acoustic parameterization. For the face, space coordinates of key-points, e.g. around the mouth, could be used or alternatively parameters based on a more sophisticated face model.

For the maximization, the distribution $p(\mathbf{y})$ is irrelevant since it does not depend on $\mathbf{x}$. Distribution $p(\mathbf{x}) \sim N(\mathbf{x}; \bar{\mathbf{x}}, \sigma_x)$ is assumed to be Gaussian, for simplicity. The relationship between the AV and articulatory parameter vectors is in general expected to be nonlinear but could be to a first order stochastically approximated by a linear mapping (both $\mathbf{x}$ and $\mathbf{y}$ are centered by mean subtraction):

$$\mathbf{y} = W\mathbf{x} + \epsilon \quad (2)$$

The error ϵ of the approximation is regarded as zero-mean Gaussian with covariance Q. The maximum a posteriori probability solution is:

$$\hat{\mathbf{x}} = (\sigma_x^{-1} + W^T Q^{-1} W)^{-1}(\sigma_x^{-1}\bar{\mathbf{x}} + W^T Q^{-1}\mathbf{y}) \quad (3)$$

The estimated solution is a weighted mean of both the observation and the prior models. The weights are proportional to the relative reliability of the two summands.

The linear mapping can be determined by means of multivariate linear analysis techniques. Such techniques constitute a class of well studied methods in statistics and other quantitative disciplines; one can find a comprehensive introduction in [9]. We can easily see that, when we completely know the underlying second-order statistics in the form of covariance matrices R_{xx}, R_{yy}, and R_{yx}, then the optimal in the MSE sense choice for the $m \times n$ matrix W corresponds to the Wiener filter

$$W = R_{yx}R_{xx}^{-1}, \quad (4)$$

and the covariance of the approximation error in (2) is $Q \triangleq E\{(y - \hat{y})(y - \hat{y})^T\} = R_{yy} - R_{yx}R_{xx}^{-1}R_{yx}^T$.

Since the second order statistics are in practice unknown a-priori, we must contend ourselves with sample-based estimates thereof; for example, if the $N \times n$ matrix X gathers N samples of $\mathbf{x}$, then a reasonable estimate is $R_{xx} \approx \frac{1}{N}X^T X$, and similarly for R_{yy}, and R_{yx}. These estimates may not be reliable enough when the training set size N is small relatively to the feature dimensions n of $\mathbf{x}$, m of $\mathbf{y}$, and, consequently, when plugged into (4) to yield W, can lead to quite poor performance when we apply the linear regressor (2) to unknown data. This is the main reason why in [7] we proposed the application of Canonical Correlation Analysis (CCA) to estimate the linear mapping. Among other benefits, we saw that CCA provides a sound mechanism to select reduced-rank multivariate linear regression models which can outperform the conventional full-rank model in the small training set size case.

2.2. Determination of Articulatory Parameter Trajectories

This framework can be extended to handle the inversion of time-varying AV parameter sequences. The probabilities in Eq. (1) will now concern vector sequences. The main consideration is to find accurate observation and prior models that would make the solution tractable. This is not straightforward given the complexity of the relationship between the acoustic and the articulatory space, which in general is nonlinear and one-to-many. Further, visual information should be properly exploited in order to somehow constrain inversion and reduce the number of possible solutions. Motivated by current research in AV speech recognition, in [7] we extended the work in [6] to multistream HMMs in order to better fuse the audio and visual modalities.

Intuitively, in the case of continuous speech, we expect the linear approximation of Eq. (2) to only be acceptable for limited time intervals corresponding to a specific phoneme, or at least a part of the phoneme. We also expect that using different, phoneme-specific mappings would be even more effective. Hence, we would have a piecewise linear approximation for the observation model. As a prior model for the dynamics of the articulatory parameters, an HMM is used. Articulator dynamics are in general expected to be phoneme-dependent and so we have one HMM for each phoneme and one articulatory-to-audiovisual mapping for each state. Further, as in audiovisual speech recognition [10] we assume that the audio and visual cues form two separate streams y_a and y_v correspondingly which are weighted differently when determining the HMM c output probability $p(c|\mathbf{y}) \propto N(\mathbf{y_a}; \mathbf{m_{c,a}}, \Sigma_{c,a})^{w_a} N(\mathbf{y_v}; \mathbf{m_{c,v}}\Sigma_{c,\text{v}})^{w_\text{v}}$. We accept that the weights w_a and w_ν should sum to one. The distribution of the articulatory parameters at each HMM state is Gaussian. A separate linear mapping $\mathbf{y} = W_j\mathbf{x} + \epsilon_j$ is considered at each state.

Speech inversion involves finding the optimal state sequence given the audiovisual data and then for each state-aligned analysis frame estimate the corresponding articulatory parameters as in Eq.3, exploiting the state-specific linear mapping. The state sequence is found by the Viterbi algorithm using the audiovisual data in two properly weighted streams. *HMM training* is performed in the conventional way by likelihood maximization [6]. Given the occupation probabilities at each state, the linear mappings between audiovisual and articulatory data are estimated by means of reduced-rank canonical correlation analysis.

Alternatively, to account for possible asynchrony between the involved modalities, we can model their dynamics by using separate audio and visual HMMs. Having an estimate of the articulatory trajectories based on each modality late fusion is then possible. The final predicted trajectories can be generated as a weighted average. In case of independence it is straightforward to derive the corresponding weights, by properly adapting (3). For improved efficiency, viseme- instead of phoneme-level HMMs may be used for the visual stream. Visemes correspond to groups of phonemes that are indis-

tinguishable from each other when viewed on the face. For example, the viseme *P* corresponds to the group of phonemes /p/,/b/,/m/. The visemes that have been used in the experiments are given in [11]. We will see that this modeling scheme may lead to improved performance.

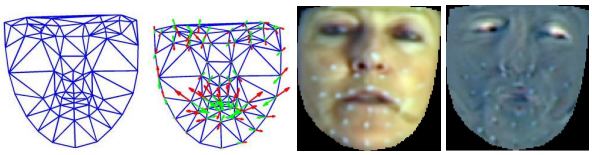

Fig. 2. Visual Front-End. *Left*: Mean shape s_0 and the first eigenshape s_1. *Right*: Mean texture A_0 and the first eigentexture A_1.

2.3. Face Active Appearance Modeling

We use *Active Appearance Models* (AAM) [12] of faces to accurately track the speaker's face and extract visual speech features from both its shape and texture. AAM are generative models of object appearance and have proven particularly effective in modeling human faces for diverse applications, such as face recognition or tracking. In the AAM scheme an object's shape is modeled as a wireframe mask defined by a set of landmark points $\{x_i, i = 1 \dots N\}$, whose coordinates constitute a shape vector s of length $2N$. We allow for deviations from the mean shape s_0 by letting s lie in a linear n-dimensional subspace, yielding $s = s_0 + \sum_{i=1}^{n} p_i s_i$. The deformation of the shape s to the mean shape s_0 defines a mapping $W(x;p)$, which brings the face exemplar on the current frame I into registration with the mean face template. After canceling out shape deformation, the face color texture registered with the mean face can be modeled as a weighted sum of "eigentextures" $\{A_i\}$, *i.e.*, $I(W(x;p)) \approx A_0(x) + \sum_{i=1}^{m} \lambda_i A_i(x)$, where A_0 is the mean texture of faces. Both eigenshape and eigentexture bases are learned during a training phase.The first few of them extracted by such a procedure are depicted in Fig. 2.

Given a trained AAM, model fitting amounts to finding for each video frame I_t the parameters $\tilde{p}_t \equiv \{p_t, \lambda_t\}$ which minimize the squared texture reconstruction error $I_t(W(p_t)) - A_0 - \sum_{i=1}^{m} \lambda_{t,i} A_i$; efficient iterative algorithms for this non-linear least squares problem can be found in [12]. The fitting procedure employs a face detector [13] to get an initial shape estimate for the first frame. As visual feature vector for speech inversion we use the parameters $\tilde{p}_t$ of the fitted AAM.

3. EXPERIMENTS AND DISCUSSION

Database Description For our experiments we have used the QSMT dataset described in detail in [8]. This dataset contains simultaneous measurements of the audio signal, tongue movements and facial motion during speech. In short, apart from the audio signal which is sampled at 16kHz and the video which is at 30fps, each frame of the dataset (at the rate of 60 fps) contains the 3D coordinates of 25 reflectors glued on the speaker's face (Qualisys/QS data, 75-dimensional vector $\mathbf{x}$), as well the 2D mid-sagittal plane coordinates of 6 EMA (Electromagnetic Articulography) coils glued on the speaker's tongue, teeth and lips (12-dimensional vector $\mathbf{y}$), comprising in total around 65000 data pairs (x_t, y_t). These correspond to one repetition of 135 symmetric VCV (Vowel-Consonant-Vowel) and 37 CVC (Consonant-Vowel-Consonant) words and 266 short everyday Swedish sentences. All data are temporally aligned and phoneme-level transcriptions are included as well. The data aqcuisition setup is shown in Fig. 1.

Next, we give our experiments in audiovisual speech inversion. To represent the speech signal we use 16 MFCCs (A). They are extracted from 35-ms preemphasized (coefficient: 0.97) and Hamming windowed frames of the signal, at 60Hz, to match the frame rate at which the EMA data are recorded. The 0-th coefficient is excluded. On the articulatory side, we use the 2D coordinates of the 3 coils on the tongue (tip, blade, dorsum) and the coil on the lower incisor. The data have been centered by mean subtraction. For the face, after active apperance modeling, we have utilized 7 features representing shape and 17 representing apperance, i.e. 24 parameters (AAM) in total. Alternatively, for comparison and to also show the full potential of utilizing facial information for inversion, all the 3D coordinates of the face markers have been used as they are provided in the database, i.e. 75 features (QS).

We have built models to recover articulatory trajectories either from acoustic (A) and facial data (AAM, QS) separately or from both combined (A-AAM, A-QS). To investigate the incorporation of visual information via AAM we could only use a subset of the QSMT database corresponding to all the VCV sequences and half of the Swedish sentences for which video of sufficiently good quality was available. For training, we have randomly selected 90% of these utterances and testing is performed on the rest 10%.

To evaluate the obtained results we have estimated both the RMS difference between the original x and the estimated $\hat{x}$ trajectories as well as the Pearson product-moment correlation coefficient, $\rho_{x\hat{x}} = tr(E[x\hat{x}^T])/\sqrt{tr(E[xx^T])tr(E[\hat{x}\hat{x}^T])}$. Results are summarized in Fig. 3. The correlation coefficient and the RMS error for the predicted trajectories are shown for increasing number of HMM states. One left-right HMM per phoneme and one separate for silence have been trained. The results at zero states correspond to global linear models and are included for comparison.

For the audiovisual case (A-AAM, A-QS) multistream HMMs have been used. The stream weights are essentially applied only for the determination of the optimal HMM state sequence via the Viterbi algorithm. This process is actually an alignment and not a recognition process, as we consider that the phonemic content of each utterance is known. We have found that the performance is optimal in case the alignment is performed using only the audio features, that is if we assign a zero stream weight to the visual stream. This observation is in accordance with similar experience in audiovisual speech recognition for audio-noise free experiments [10]. The audio should be exclusively trusted for recognition when no noise is present. In our audiovisual-to-articulatory inversion setup it appears that in the absence of audio noise, the audio stream should be trusted for alignment but, given the optimal state alignment, the contribution of the visual modality in inversion is very important in any case.

In general, fusion of the visual AAM features with audio (A-AAM) is beneficial compared to the audio-only (A) or visual-only (V) cases. Of course, the best performance is achieved when audio is fused with the ground-truth facial features (QS). This is justified since the latter accurately represent 3D facial information, which is clearly richer than the 2D image based information captured from the AAM features. Measurement of the AAM features however is much more practical since it does not require any special or inconvenient acquisition setup but only images from the frontal view of the speaker's face.

At a different level, we have explored various modeling and fusion schemes of the audio and visual stream dynamics in the proposed framework. This time our experiments were performed on the full QSMT dataset and the visual information was represented by the

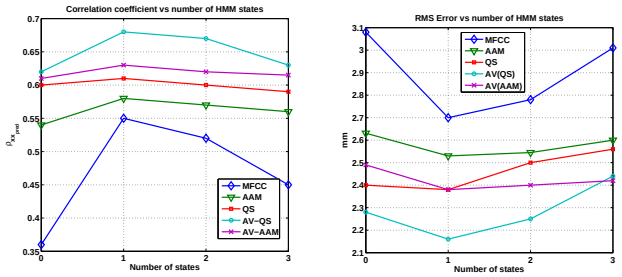

Fig. 3. Correlation coefficient and RMS Error between original and predicted articulatory trajectories for increasing number of HMM states using facial information only (via AAM or Qualisys (QS) features) audio only (MFCC) and both (AV-AAM, AV-QS). Zero states correspond to the case of a global linear model.

Features	Level	Type	States	RMS (mm)	$\rho_{x\hat{x}}$
Audio	P	HMM	2	2.56	0.60
QS	P	HMM	2	2.30	0.65
QS	V	HMM	3	2.24	0.66
A-QS	P	HMM	2	2.16	0.69
A-QS	P-P	HMM+LF	2-2	2.02	0.71
A-QS	P-V	HMM+LF	2-2	1.99	0.72
A-QS	**P**	**MS-HMM**	**2**	**1.95**	**0.74**

Table 1. RMS error and correlation coefficient for the predicted articulatory trajectories using various HMM-based schemes. Audio features (A), 3D facial marker coordinates tracked from Qualisys (QS) or both have been used. The models may be either at the phoneme (P) or at the viseme (V) level and either single HMMs are used, or in a late fusion (LF) configuration or as multistream (MS-HMM).

Qualisys features (QS). Results are given in Table 1. Audio and visual information dynamics have been fused in three different ways, namely via simple HMMs trained on concatenated feature vectors, single-per modality HMMs with late fusion as sketched in subsection 2.2 (HMM+LF), or finally via multistream HMMs (MS-HMM). For the late fusion scheme two variants are given, differentiating from each other in whether the visual stream is modeled as a sequence of phonemes (P) or visemes (V). Interestingly, the visemes demonstrate improved performance, both in the single modality case and in fusion.

An example of the predicted trajectories for the 2D coordinates of the tongue blade on the midsagittal plane against the measured ones is shown in Fig.4 for a Swedish phrase. The corresponding RMS error is 2.13 mm for the multistream HMM case trained on MFCC and Qualisys feature sets.

4. CONCLUSIONS AND FUTURE WORK

We have elaborated on a framework based on Hidden Markov Models to perform audiovisual-to-articulatory speech inversion. Experiments have been carried out on the QSMT dataset to recover EMA coil movements from face motion and speech acoustics. Face is modeled by means of Active Appearance Modeling. In this way it is possible to utilize visual information without a special acquisition setup as the Qualisys system, that would require for example gluing markers on the speaker's face. Performance may slightly degrade compared to the case when these markers are used but it

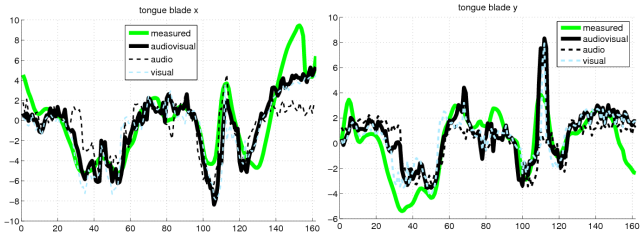

Fig. 4. Coordinates of the coil on the tongue blade as predicted from audio only (A), face only (Qualisys) and both (A-QS MS-HMM). The measured coordinates are also superimposed with light colored thick lines.

is clearly seen that, in the audiovisual case, inversion is still possible with satisfactory performance and clearly outpeforms the corresponding single-modality cases. Experiments regarding modeling and fusion schemes additionaly show that modeling the visual stream at the viseme level may improve performance and that the MS-HMM outperforms other rival schemes, such as the use of separate HMMs and late fusion. Currently, we have also been exploring the use of Product-HMMs that could further improve performance [14]. These could account for asynchrony as the late fusion scheme does but in a more constrained and robust manner. We further look into modifications concerning continuity and more detailed imposition of dynamic constraints, e.g. related to coarticulation. In parallel, a more detailed phoneme/viseme-based analysis is under way and is expected to unveil the full benefits of the proposed framework.

5. REFERENCES

[1] H. Kjellstrom, O. Engwall, and O. Balter, "Reconstructing tongue movements from audio and video," in *Interspeech*, 2006, pp. 2238–2241.

[2] O. Engwall, "Introducing visual cues in acoustic-to-articulatory inversion," in *INTERSPEECH*, 2005, pp. 3205–3208.

[3] J. Jiang, A. Alwan, P. A. Keating, E. T. Auer Jr., and L. E. Bernstein, "On the relationship between face movements, tongue movements, and speech acoustics," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1174–1188, 2002.

[4] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Sp. Comm.*, vol. 26, pp. 23–43, 1998.

[5] K. Richmond, S. King, and P. Taylor, "Modelling the uncertainty in recovering articulation from acoustics," *Computer Speech and Language*, vol. 17, pp. 153–172, 2003.

[6] S. Hiroya and M. Honda, "Estimation of articulatory movements from speech acoustics using an hmm-based speech production model," *IEEE TSAP*, vol. 12, no. 2, pp. 175–185, March 2004.

[7] A. Katsamanis, G. Papandreou, and P. Maragos, "Audiovisual-to-articulatory speech inversion using hmms," in *Proceedings of IEEE Int'l Workshop on Multimedia Signal Processing (MMSP 2007)*.

[8] O. Engwall and J. Beskow, "Resynthesis of 3d tongue movements from facial data," in *EUROSPEECH*, 2003.

[9] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Acad. Press, 1979.

[10] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Tr. Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[11] J. Beskow, "Rule-based visual speech synthesis," in *Proc. of the 4th European Conference on Speech Communication and Technology (EUROSPEECH 95)*, 1995.

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[13] P. Viola and M.J. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. on Comp. Vision and Pat. Recog.*, 2001, vol. I, pp. 511–518.

[14] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," in *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, 2001.