FRAME-DEPENDENT MULTI-STREAM RELIABILITY INDICATORS FOR AUDIO-VISUAL SPEECH RECOGNITION

Ashutosh Garg,[*] *Gerasimos Potamianos,*[+] *Chalapathy Neti,*[+] and *Thomas S. Huang*[*]

[+] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
[*] Beckman Institute, University of Illinois, Urbana, IL 61801, USA
E-mails: [+] {gpotam,cneti}@us.ibm.com; [*] {ashutosh,huang}@ifp.uiuc.edu

ABSTRACT

We investigate the use of local, frame-dependent reliability indicators of the audio and visual modalities, as a means of estimating stream exponents of multi-stream hidden Markov models for audio-visual automatic speech recognition. We consider two such indicators at each modality, defined as functions of the speech-class conditional observation probabilities of appropriate audio- or visual-only classifiers. We subsequently map the four reliability indicators into the stream exponents of a state-synchronous, two-stream hidden Markov model, as a sigmoid function of their linear combination. We propose two algorithms to estimate the sigmoid weights, based on the maximum conditional likelihood and minimum classification error criteria. We demonstrate the superiority of the proposed approach on a connected-digit audio-visual speech recognition task, under varying audio channel noise conditions. Indeed, the use of the estimated, frame-dependent stream exponents results in a significantly smaller word error rate than using global stream exponents. In addition, it outperforms utterance-level exponents, even though the latter utilize a-priori knowledge of the utterance noise level.

1. INTRODUCTION

Automatic speech recognition (ASR) using information from the video of the speaker's face, in addition to the traditional audio, has been an active area of research in recent years [1]-[6]. Such work has been well motivated by human speech perception [7], as well as by the obvious visual signal robustness to acoustic degradation.

A challenging problem in audio-visual ASR is the integration (fusion) of the two heterogeneous sources of speech information [1]. A number of techniques have been proposed in the literature for this task. It is generally agreed that the combination of single-modality (audio- and visual-only) classifier outputs (e.g., observation likelihoods), also known as the *decision fusion* framework, outperforms fusion at the feature level [3]-[6]. Typically, decision fusion linearly combines the two classifier scores, where the classifiers can be neural networks [6], [8], or, more commonly, *hidden Markov models* (HMMs) [3]-[5], and the integration level can vary, allowing for example audio-visual asynchrony, as is the case in the product [3], [4] and coupled HMMs [5]. The linear combination weights manipulate the contribution of each modality to the recognition process, hopefully capturing the reliability of the audio and visual observation streams.

In the literature, such combination weights have been usually set to *global* values over an entire dataset, either constant for all classes of interest (such as HMM states) [3]-[5], or dependent on the class labels [4]. Occasionaly, they have been chosen to be utterance-dependent, based on estimates of the audio signal (using the "voicing index" [4], or the signal-to-noise ratio [8], [9]), and utterance- [9], [10], or frame-dependent [6], based on confidence estimates of the audio-only classifier. However, in practical audio-visual ASR, the speech information carried by *both* the audio and video signals can vary dramatically, and at a very *local*, temporal level. For example, possible noise bursts, face occlusion, or face tracking failures can greatly change the reliability of the affected stream in ASR. Clearly, frame-dependent combination weights that capture the information content of both modalities are needed to handle this scenario.

In this paper, we propose an algorithm to estimate such frame-dependent weights within the framework of the state-synchronous multi-stream HMM, a widely used model for audio-visual decision fusion [3]-[5]. The algorithm utilizes well known indicators of the confidence of both audio- and visual-only classifiers [9], [11] to capture the reliability of the two streams of interest, and then estimates a sigmoid, that maps their values to the desirable decision fusion weights. Both the proposed mapping and the estimation algorithm constitute the contributions of this work.

The remaining of the paper is structured as follows: Section 2 reviews the multi-stream HMM framework for audio-visual speech recognition, and Section 3 introduces the reliability indicators used. Section 4 is devoted to the audio-visual combination weight estimation algorithm, based on the frame-dependent modality reliability indicators. Section 5 describes the audio-visual database and ASR experiments, and finally, Section 6 summarizes the paper.

2. THE MULTI-STREAM HMM

The main concentration of this paper is modeling the reliability of the two modalities in decision fusion for audio-visual ASR. Since temporal modeling in not our focus, we restrict ourselves to the popular, state-synchronous *multi-stream HMM* (MSHMM) as the statistical model for audio-visual integration [3], [4]. Extensions of our proposed algorithm to models that allow audio-visual state asynchrony, such as the product [3] and coupled HMMs [5], or to neural network classifiers [6], can easily be devised.

The MSHMM is a variant of the standard HMM [12], where instead of a single observation stream, there exist multiple streams of information (in our case, two: one for each modality, audio and visual). Given a time-synchronous, bimodal (audio-visual) observation vector $\mathbf{o}_t = [\, \mathbf{o}_{a,t}\, ,\, \mathbf{o}_{v,t}\,]$ at time instant t ("frame"), the MSHMM models its class-conditional likelihood as the product of the observation likelihoods of its single-stream components, raised
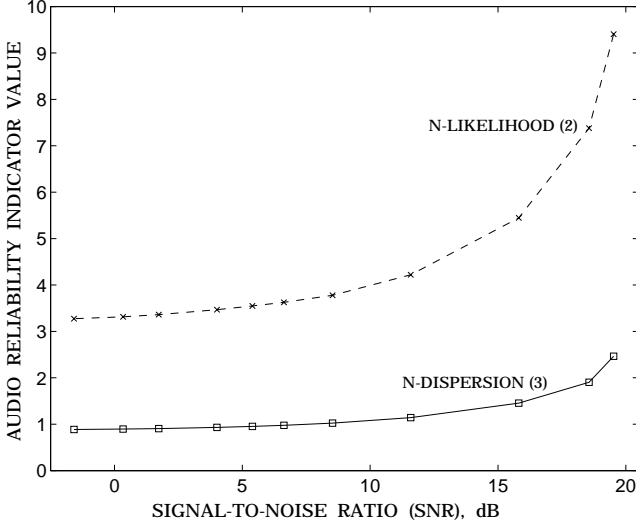
Fig. 1. Audio reliability indicators $\mathcal{L}_{a,t}$ and $\mathcal{D}_{a,t}$ (see (2) and (3)), depicted as a function of the noise level, present in the audio data.

to appropriate *stream exponents*, namely

$$P\left(\mathbf{o}_t \mid c\right) = \prod_{s \in \{a,v\}} P\left(\mathbf{o}_{s,t} \mid c\right)^{\lambda_{s,t}} . \qquad (1)$$

In (1), $c \in C$ denote the speech classes of interest (such as context-dependent, sub-phonetic units), and $P(\mathbf{o}_{s,t}|c)$, $s = a, v$, are the audio- and visual-only emission probabilities, typically considered to be *Gaussian mixture models* (GMMs). Stream exponents $\lambda_{s,t}$ are in general non-negative, and, in this work, they are also assumed to add to one, i.e., $\lambda_{a,t} + \lambda_{v,t} = 1$, for all t. Notice that, due to their presence, (1) does not represent a probability density function. Instead, it can be thought of as a *scoring* function. This viewpoint allows us to analyze it similarly to the standard maximum likelihood framework, and employ the *expectation-maximization* (EM) algorithm [12] to estimate the MSHMM parameters, where the expectation step can be performed separately for each GMM, or jointly for the entire model.

Exponents $\lambda_{s,t}$ provide a means to model the reliability of each feature stream, by allowing one to manipulate the contribution of each modality to the recognition score. In realistic audio-visual ASR, such reliabilities can rapidly vary at a temporal level. Therefore, in this work, we consider the exponents to be time varying, defined as a function of local (frame-dependent) reliability indicators of the two streams of interest. As discussed in the following section, such stream reliability indicators depend on the corresponding local modality observations, and the respective statistical model that is assumed to generate them. As a result, the stream exponents become a function of the *joint* audio-visual feature vector $\mathbf{o}_t$. Thus, (1) differs to work reported elsewhere, where exponents are set to "global" weights, constant over a whole dataset [3]-[5], possibly depending on the class label c [4], or locally varying, but depending only on the audio observation $\mathbf{o}_{a,t}$ [4], [6], [9], [10].

3. STREAM RELIABILITY INDICATORS

A number of functions have been proposed in the literature as a means of assessing the reliability of the class information that is

Reliability Indicator	Correlation with audio-only WER	Correlation with visual-only WER
$\mathcal{L}_a$	-0.7434	0.0183
$\mathcal{L}_v$	0.1041	-0.2191
$\mathcal{D}_a$	-0.7589	0.0126
$\mathcal{D}_v$	0.1014	-0.2066

Table 1. Correlation of the two stream reliability indicators (2) and (3) with the audio- and visual-only word error rates (WERs).

contained in an observation, assumed to be modeled by a particular classifier [9]-[11]. Following prior work [11], we select two reliability indicators for each of the two streams. Given the stream observation $\mathbf{o}_{s,t}$, both indicators utilize the class-conditional observation likelihoods of its N-best most likely generative classes, denoted by $c_{s,t,n} \in C$, $n = 1,...,N$. These are ranked according to descending values of $P(\mathbf{o}_{s,t}|c)$, $c \in C$ (see also (1)).

The first reliability indicator is the N-best log-likelihood difference, defined as

$$\mathcal{L}_{s,t} = \frac{1}{N-1} \sum_{n=2}^{N} \log \frac{P\left(\mathbf{o}_{s,t} \mid c_{s,t,1}\right)}{P\left(\mathbf{o}_{s,t} \mid c_{s,t,n}\right)} . \qquad (2)$$

This is chosen, since it is argued that the likelihood ratios between the first N classification decisions are informative about the class discrimination. The second selected reliability indicator is the N-best log-likelihood dispersion. This is defined as

$$\mathcal{D}_{s,t} = \frac{2}{N(N-1)} \sum_{n=1}^{N} \sum_{n'=n+1}^{N} \log \frac{P\left(\mathbf{o}_{s,t} \mid c_{s,t,n}\right)}{P\left(\mathbf{o}_{s,t} \mid c_{s,t,n'}\right)} . \qquad (3)$$

The main advantage of (3) over (2) lies on the fact that (3) captures additional N-best class likelihood ratios, not present in (2). In our analysis, we choose N to be 5.

In the remainder of this section, we argue that the selected indicators do capture the reliability of the speech class information, available in the stream of interest. For example, such information, on basis of the audio channel alone, is expected to degrade, as the audio becomes corrupted by increasing levels of noise. Fig. 1 demonstrates that both $\mathcal{L}_{a,t}$ and $\mathcal{D}_{a,t}$ successfully convey the degradation of the audio stream reliability, since they are monotonic on the signal-to-noise ratio (see Section 5 for the data and the experiment design).

Of course, our primary interest lies in minimizing the word error rate based on MSHMM (1). To further justify the selection of the four reliability indicators in audio-visual ASR, we report a *correlation* analysis between the values of these indicators, averaged at the utterance level, and the utterance *word error rate* (WER). The results are summarized in Table 1, and they clearly demonstrate the presence of significant within-stream correlation. As expected, there is low correlation across streams. These observations argue favorably for using reliability indicators of both audio and visual streams in audio-visual ASR.

4. RELIABILITY BASED STREAM EXPONENTS

4.1. Reliability Indicator to Stream Exponent Mapping

We would like now to estimate a mapping from the four selected reliability indicators to the desired MSHMM stream exponent $\lambda_{a,t}$,

and its derived $\lambda_{v,t} = 1 - \lambda_{a,t}$. We choose to use a sigmoid function for this task, due to its nice properties: The sigmoid is bounded within zero and one, and it is monotonic, and smooth. For simplicity, let us denote by $\mathbf{d}_t$ the vector of the four selected indicators, namely $[d_{1,t}, d_{2,t}, d_{3,t}, d_{4,t}] = [\mathcal{L}_{a,t}, \mathcal{L}_{v,t}, \mathcal{D}_{a,t}, \mathcal{D}_{v,t}]$. Then, the sigmoid mapping is defined as

$$\lambda_{a,t} = \frac{1}{1 + \exp\left(-\sum_{i=1}^{4} \mathrm{w}_i \, d_{i,t}\right)}, \qquad (4)$$

where $\mathbf{w} = [\mathrm{w}_1, \mathrm{w}_2, \mathrm{w}_3, \mathrm{w}_4]$ is the vector of the mapping parameters. In the following, we propose two algorithms to estimate $\mathbf{w}$, given frame-level labeled audio-visual observations $\{(\mathbf{o}_t, c_t), t \in \mathcal{T}\}$, for a training set of time instants $\mathcal{T}$. The first algorithm seeks maximum conditional likelihood estimates of parameters $\mathbf{w}$, under the MSHMM observation model (1), whereas the second method seeks a $\mathbf{w}$ that minimizes the misclassification error on set $\mathcal{T}$. Notice that the required training set labels c_t, $t \in \mathcal{T}$, can be obtained by a forced alignment of the training set utterances using a suitable HMM.

4.2. Maximum Conditional Likelihood Parameter Estimation

Given an audio-visual observation vector $\mathbf{o}_t$, we represent the conditional likelihood of the class $c \in C$ by

$$P(c \mid \mathbf{o}_t) = \frac{P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}}{\sum_{c \in C} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}}, \qquad (5)$$

under the assumption of a uniform class prior $P(c)$ (see also (1)).

We then seek parameters $\mathbf{w}$ of (4) that result to *maximum conditional likelihood* (MCL) of the training data labels, namely

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{t \in \mathcal{T}} \log P(c_t \mid \mathbf{o}_t). \qquad (6)$$

The above optimization problem is solved iteratively, by performing a gradient descent on (6), with respect to $\mathbf{w}$, as

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \eta \sum_{t \in \mathcal{T}} \left.\frac{\partial \log P(c_t|\mathbf{o}_t)}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}^{(k)}}, \qquad (7)$$

for $k = 0, 1, 2, \ldots$, where the gradient vector elements are given by

$$\frac{\partial \log P(c_t|\mathbf{o}_t)}{\partial \mathrm{w}_i} = \lambda_{a,t}(1 - \lambda_{a,t}) \, d_{i,t} \left[\log \frac{P(\mathbf{o}_{a,t}|c_t)}{P(\mathbf{o}_{v,t}|c_t)} \right.$$

$$\left. - \frac{\sum_{c \in C} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}} \log \frac{P(\mathbf{o}_{a,t}|c)}{P(\mathbf{o}_{v,t}|c)}}{\sum_{c \in C} P(\mathbf{o}_{a,t}|c)^{\lambda_{a,t}} P(\mathbf{o}_{v,t}|c)^{1-\lambda_{a,t}}} \right],$$

for $1 \le i \le 4$ (see also (4)-(6)). In (7), we choose $\mathbf{w}^{(0)} = [1,1,1,1]$. The learning rate parameter η controls convergence speed, and since (6) is not a convex optimization problem, η needs to be kept relatively small. In our experiments, when choosing $\eta = 0.01$, convergence is typically achieved within a few tens of iterations.

4.3. Minimum Classification Error Parameter Estimation

The second technique adopted in this work for the estimation of the sigmoid parameters $\mathbf{w}$ is the *minimum classification error* (MCE) approach. Here, instead of maximizing the conditional likelihood, we need to perform a grid search over the parameter space, and

choose the parameter vector $\hat{\mathbf{w}}$ that maximizes the frame level classification performance on the training set $\mathcal{T}$. In this particular task, since four reliability indicators are selected, we need to compute the value of four parameters. A grid search over such a parameter space is not impossible, however since each weight w_i can vary from $-\infty$ to ∞, it cannot be carried out exactly.

To simplify the search problem, we make use of the MCL parameter estimates of $\mathbf{w}$, obtained as discussed in the previous subsection, in order to obtain the approximate dynamic range for the parameters and limit the search within it. Then, for each parameter vector value over the reduced grid, we compute the frame error. The weight assignment that results in the best performance (minimum classification error at the frame level) is chosen as the output.

5. DATABASE AND EXPERIMENTS

To validate the performance of the proposed scheme, we conduct audio-visual ASR experiments on a multi-subject, connected-digit recognition task. The database consists of synchronously captured, high-quality audio and video of 50 subjects, uttering 7- or 10-tuple strings of connected digits. Approximately 10 hrs of such data are available. Details of this database can be found in [13].

Time-synchronous audio and visual features are extracted from this database using the algorithms reported in [4]. Briefly, the audio stream features are obtained by a linear discriminant analysis (LDA) feature projection, applied on the concatenation of neighboring audio mel-frequency cepstral coefficient vectors, followed by a maximum likelihood linear transform (MLLT). The visual features are obtained by an LDA/MLLT cascade, applied on the discrete cosine transform coefficients of the pixel values of a properly normalized region-of-interest (ROI). Such ROI contains the mouth and jaw area of the subject, detected by means of a statistical face tracking algorithm. The visual features are upsampled to the audio feature extraction rate (100 Hz) using linear interpolation, thus allowing audio-visual speech modeling with the MSHMM (1). All feature transform matrices are estimated on the training part of the database.

For the 11-word digit vocabulary (includes both "zero" and "oh"), a set of 22 phones are used in ASR, with 104 context-dependent sub-phonetic HMM states, and approximately 5.3k Gaussian mixture components per stream. The parameters of both audio- and visual-only HMMs are separately estimated by EM on the training part of the database, and the two models joined into the MSHMM (1).

To test the performance of our algorithm over varying stream reliability conditions, we artificially add speech babble noise to the database audio, at various levels of *signal-to-noise ratio* (SNR). Such noise is added to the test and held-out sets of the database only, thus creating a mismatch to the audio-only HMMs, that are trained on the original clean database audio, as mentioned above.

In Table 2, we summarize our experimental results. We report both *frame misclassification error* (FER - as compared to the 22 phone labels of the forced alignment of the test set clean audio data), as well as the ASR *word error rate* (WER), %. We first consider the audio-only and video-only performance, using the single-modality HMMs. Subsequently, we estimate a global audio exponent, constant over the entire dataset and all classes, that minimizes the MSHMM based audio-visual WER on the held-out set. Such exponent λ_a is estimated after a grid search at the resolution of 0.01. Not surprisingly, audio-visual ASR significantly outperforms both audio- and visual-only WERs, demonstrating the

Condition	FER	WER
Audio-Only	58.80	30.29
Visual-Only	41.18	20.45
AV-Global	31.80	10.35
AV-Frame, MCL	31.53	10.13
AV-Frame, MCE	31.18	8.64

Table 2. Frame misclassification (FER) and word error rates (WER), %, for multi-stream HMM based audio-visual digit recognition using global vs. frame dependent exponents, estimated by means of the proposed algorithm. Audio-only and visual-only recognition results are also depicted. Noise at a number of SNRs has been added to the audio utterances.

suitability of MSHMM based decision fusion. We then use the four reliability measures of Section 3 to obtain frame-dependent stream exponents by means of (4). Both MCL and MCE algorithms, introduced in Section 4, are employed to estimate the regression parameter vector $\mathbf{w}$. Both approaches further reduce FER, as well as the WER, with the MCE based estimation resulting in a 17% relative WER reduction, over the use of global weights. It is interesting to compare these WERs to the scenario that uses utterance-dependent exponents, and assumes a-priori knowledge of the SNR (a best case scenario for SNR-dependent exponent estimation). Such exponents are estimated on held-out data matched to the noise level, and are depicted in Fig. 2. Even in this "cheating" case, the resulting 9.08% WER is worse than the WER achieved by frame-dependent exponents with MCE estimation of the parameters.

The results presented above clearly demonstrate the superiority of our approach over existing schemes for weighting the different modalities for audio-visual ASR. It significantly outperforms the use of global optimal weights, and, interestingly, MCE based sigmoid parameter estimation even beats the "cheating" case of SNR-dependent exponents with known noise degradation level. We believe that this occurs because our approach captures both audio and video stream reliabilities, and jointly uses them to estimate the audio-visual fusion exponents.

6. CONCLUSIONS

We considered stream reliability measures for estimating adaptive, frame-dependent decision fusion weights for improved audio-visual speech recognition by means of multi-stream HMMs. We proposed a novel *reliability to fusion weight mapping* and presented two estimation algorithms (MCL and MCE) of the mapping parameters. This paper thus extended previous work, where the fusion weights were either limited to constant, or audio-noise dependent values. The reported recognition results demonstrated the superiority of the introduced technique.

7. REFERENCES

[1] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Humans and Machines*. Berlin: Springer, pp. 331–349, 1996.

[2] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Sig. Process. Mag., 18*(1):9–21, 2001.
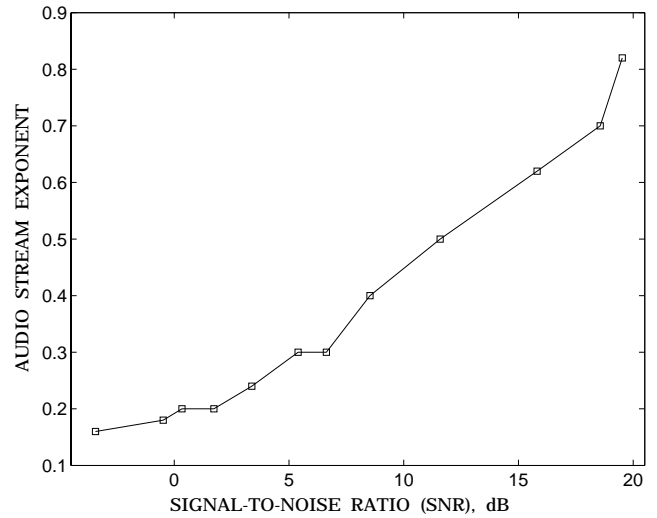

Fig. 2. Optimal MSHMM audio stream exponent as a function of the audio channel SNR. The audio-only MSHMM component has been trained on clean audio data (20 dB SNR).

[3] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multim., 2*(3):141–151, 2000.

[4] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D.Vergyri, "Large-vocabulary audio-visual speech recognition:A summary of the Johns Hopkins summer 2000 workshop", *Proc. Wks. Multim. Sig. Process.*, pp. 619-624, 2001.

[5] A.V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," in press: *EURASIP J. Appl. Sig. Process.*, 2002.

[6] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," in press: *EURASIP J. Appl. Sig. Process.*, 2002.

[7] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature, 264*:746–748, 1976.

[8] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading," *Proc. Int. Conf. Acous. Speech Sig. Process.*, pp. 833–836, 1996.

[9] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in D.G. Stork and M.E. Hennecke (Eds.), *Speechreading by Humans and Machines*. Berlin: Springer, pp. 461–471, 1996.

[10] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," *Proc. Europ. Tut. Wks. Audio-Visual Speech Process.*, pp. 53–56, 1997.

[11] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, pp. 746–749, 2000.

[12] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE, 77*(2):257–285, 1989.

[13] G. Potamianos and C. Neti, "Automatic speechreading of impaired speech", *Proc. Wks. Audio-Visual Speech Process.*, pp. 177-182, 2001.