USING ENTROPY AS A STREAM RELIABILITY ESTIMATE FOR AUDIO-VISUAL SPEECH RECOGNITION

Mihai Gurban and Jean-Philippe Thiran

Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
email: {mihai.gurban, jp.thiran}@epfl.ch

ABSTRACT

We present a method for dynamically integrating audio-visual information for speech recognition, based on the estimated reliability of the audio and visual streams. Our method uses an information theoretic measure, the entropy derived from the state probability distribution for each stream, as an estimate of reliability. The two modalities, audio and video, are weighted at each time instant according to their reliability. In this way, the weights vary dynamically and are able to adapt to any type of noise in each modality, and more importantly, to unexpected variations in the level of noise.

1. INTRODUCTION

Humans use visual information subconsciously to understand speech, especially in noisy conditions, but also when the audio is clean. The same integration can be performed by computers to improve the performance of speech recognition systems, when dealing with difficult audio conditions. Audio-visual speech recognition (AVSR) improves recognition rates beyond what is possible with only audio. An overview of AVSR can be found in [1].

Most approaches to AVSR use multi-stream hidden Markov models (HMMs) as the choice recognizer, as they consistently outperform uni-modal HMMs. A key issue in building such models is the method of combining scores from the different modalities, and a popular technique is performing a weighted sum of stream log-likelihoods. The method of choosing these weights is the focus of our paper.

There are many different possibilities to choose the audio-visual weights. They can be time-varying or fixed, dependent on the particular HMM state or independent of it. However, in many systems, they are simply fixed by hand for a particular environment and database [1]. The optimal value of these static weights is found by seeking the best performance on matched data. However, in a practical system, the quality of each stream can change in time, making the adjustment of the weights at utterance level or frame level a requirement. In some approaches, the weights are found based on training or held-out data. This is the case in [2], where a smooth function of the minimum classification error (MCE) is minimized. However, in this case, the training is done at the particular signal to noise ratio (SNR) at which the training data is available. This means that in different noise conditions, the system's performance will not be optimal. Another approach is to derive the stream weights from the audio channel estimated signal to noise ratio, as in [3]. The problem here is that the reliability of the video stream is not taken into account. Yet other approaches use the dispersion of the class posterior probabilities to model the stream confidences [4]. In [5], different frame-level confidence measures are investigated, among which the dispersion of the class posteriors and their entropy. The maximum class posteriors are used as reliability estimates at frame-level in [6].

Our proposed method is related to the last methods presented. We use the entropy of the state posteriors as a reliability measure at frame-level. We propose several types of mapping this entropy to stream weights. They all perform well in different stable noise conditions and even when the noise is changing in time. Our basic assumption is that the noise level and type is unknown when the system is built, so we are aiming to build a system that performs well under a wide range of noise conditions.

The paper is organized as follows. First we review the audio-visual speech recognition systems and the multi-stream HMM framework. Then we present our weight estimation methods. Finally, we present the details of our implementation and the database used, together with results and discussion.

2. AUDIO-VISUAL SPEECH RECOGNITION

In this section we briefly present the structure of an audio-visual speech recognition system. While all such systems share common traits, they can differ in three major aspects. The first one is the visual front-end; i.e., the part of the system that tracks the region of the mouth and extracts the visual features. The second one is the audio-visual integration strategy, that is, the way audio and visual information are put together in order to reach a decision about the recognized word. Finally, the type of speech recognition system can differ depending on the particular task (isolated-word recognition, continuous speech or large-vocabulary speech recognition). Our system recognizes sequences of words separated by silence, from a small-vocabulary database.

The majority of speech recognition systems use hidden Markov models [7] (HMMs) as the underlying classifiers used to represent and recognize the spoken words. Our audio-visual system also uses a particular kind of HMMs, multi-stream HMMs, which are well-suited for multimodal processing.

2.1 Visual front-end

All audio-visual speech recognition systems require the identification and tracking of the region of interest (ROI), which can be either only the mouth, or a larger region, like the entire face. This typically begins with locating the face of the speaker, using a face detection algorithm. The second step is locating the mouth of the speaker and extracting the region of interest. This region can be scaled and rotated such that

the mouth is centered and aligned.

Once the ROI has been extracted, the useful information that it contains needs to be expressed using as few features as possible. This is because the high dimensionality of the ROI impairs its accurate statistical modeling. The main types of features that can be used for visual speech recognition [1] are either appearance based features, extracted directly from the pixels of the ROI, or shape based features, extracted from the contour of the speaker's lips.

In general, the use of shape features requires a good lip tracking algorithm and makes the limiting assumption that speech information is concentrated in the contour of the lips alone. Several articles report that DCT features outperform shape based ones [8, 9]. Features can be further refined through the use of the linear discriminant transform (LDA), a transform that improve the separation between the classes, and this is our method of choice for feature extraction.

2.2 Audio-visual integration

The integration of audio and visual information [1] can be performed in several ways. The simplest one is feature concatenation [4], where the audio and video feature vectors are simply concatenated before being presented to the classifier. Here, a single classifier is trained with combined data from the two modalities.

Although the feature concatenation method of integration does lead to an improved performance, it is impossible to model the reliability of each modality, depending on the changing conditions in the audio-visual environment.

Using decision fusion, separate audio and video classifiers are trained, and their output log-likelihoods are linearly combined with appropriate weights. There are three possible levels for combining individual modality likelihoods [1]:

- Early integration, in the case when likelihoods are combined at the state level, forcing the synchrony of the two streams.
- Intermediate integration, which uses models that force synchrony at the phone or word boundaries.
- Late integration, which requires separate HMMs for each stream. The final recognized word is selected based on the n-best hypothesis of the audio and visual HMMs.

The method used in this paper is early decision fusion using a multi-stream HMM classifier [10], which will be briefly presented in the next section.

2.3 The multi-stream HMM

Multi stream HMMs are actually parallel HMMs sharing the same architecture, that is, having the same number of states and the same transitions. This forces synchrony between the modalities. The emission probability densities are modeled with Gaussian mixtures, separately for each stream, and the emission log-likelihood is computed as a weighted sum. Let $o(t) = (o_a(t), o_v(t))$ be the audio-visual feature vector (the observation) and b_{js} the corresponding likelihood arising from the Gaussian mixture for state j and stream s. Then the likelihood for stream s is [11]:

$$b_{js}(o_s(t)) = \sum_{m=1}^{M_s} c_{jsm} N(o_s(t); \mu_{jsm}, \Sigma_{jsm}) \qquad (1)$$

where $N(o; \mu, \Sigma)$ is the value in $o_s(t)$ of a multivariate gaussian with mean μ and covariance matrix Σ. M_s gaussians are used in a mixture, each weighted by c_{jsm}. The combined score b_i is then computed as:

$$\ln b_i(o(t)) = \sum_{s=a,v} \lambda_s \ln b_{is}(o_s(t)) \qquad (2)$$

This amounts to multiplying the likelihoods raised to power λ_s. The product rule is one of the most widely used probability combination rules, along with the sum rule, the min rule or the max rule [12]. These rules are compared in [13], with the purpose of combining the outputs of classifiers trained on different types of audio-only features. The product rule was found to be the best performer. The same weighted product rule can be found in [4], integrating word-level probabilities.

Typically the weights are chosen such that their sum is 1, $\lambda_a + \lambda_v = 1$, however, even in this case, b_j does not define a probability density, and should be regarded as a score. Finding the weights λ_s is the focus of our paper.

3. OUR PROPOSED METHOD

3.1 Entropy as a reliability estimate

We base our reliability estimate on the probability distribution of posterior probabilities for the HMM states Q_i. The posteriors are computed using Bayes' rule:

$$P_{is}(t) = P(Q_i|o_s(t)) = \frac{b_i(o_s(t))P(Q_i)}{\sum_j b_j(o_s(t))P(Q_j)} \qquad (3)$$

where $P(Q_i)$ is the prior probability of being in state Q_i, i being an index over all the states in all the HMMs representing the vocabulary words.

We believe that the shape of this posterior distribution is a good indicator of the reliability of its corresponding stream. If one of the posteriors is much higher than the others, there is a high probability that the classification is correct, that is, we can have a high confidence in that particular stream. In the opposite case, when the posterior distribution is flat, the confidence will be low. This idea was used in [6], where the maximum posterior probability is used to switch between streams at each time instant, choosing at each moment the most reliable one.

Our approach is to use the entropy of this posterior distribution as a reliability estimate. We compute the entropy $H_s(t)$ for stream s and time t as follows:

$$H_s(t) = -\sum_i P_{is}(t) \log_2 P_{is}(t) \qquad (4)$$

The entropy values $H_a(t)$ and $H_v(t)$ are estimates of the confidence, or rather, the lack of confidence, that we have in the respective streams. Indeed, if the entropy corresponding to one of the streams is high, it means that the posterior distribution is flat and the probability of error is high. That is, a high entropy translates into low confidence, and vice-versa. However, this intuition needs to be quantified by finding a mapping function between the entropies H_s and the corresponding stream weights λ_s. Several possibilities are explored in the following subsections.

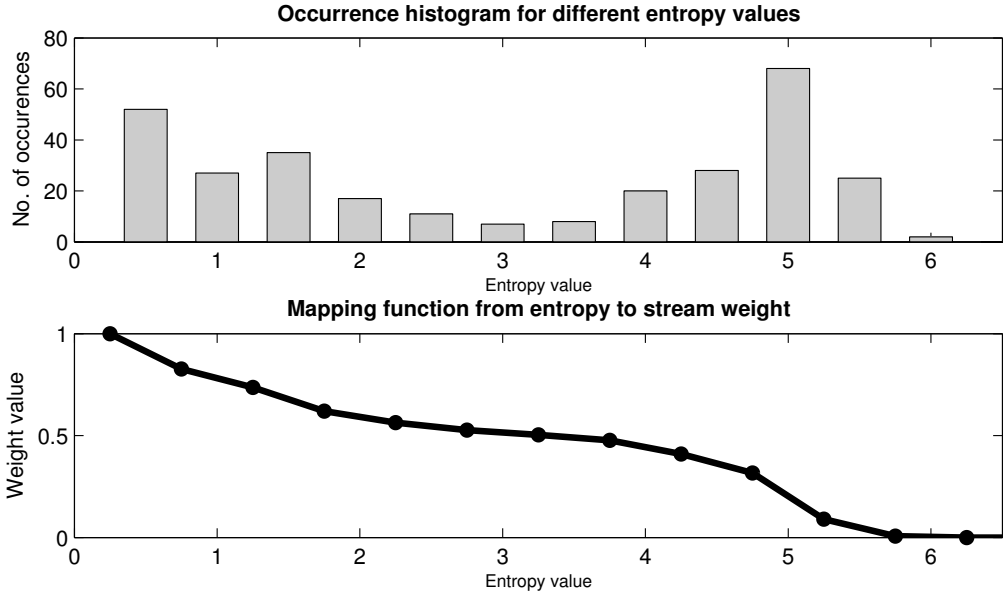

Figure 1: A flexible mapping from entropy to weight.

3.2 Negative entropy weighting

The simplest way to map entropies to weights is a linear mapping, that is

$$\lambda_s(t) = \frac{H_{max} - H_s(t)}{H_{max}} \quad (5)$$

where H_{max} is the maximum theoretical entropy for a completely flat P_{is} distribution. The weights are afterwards scaled such that their sum is 1.

3.3 Inverse entropy weighting

A simple non-linear function that can be used as a mapping is the inverse:

$$\lambda_s(t) = \frac{1/H_s(t)}{\sum_z 1/H_z(t)} \quad (6)$$

This method has been proposed for audio-only speech recognition with multiple feature streams [14]. The difference between it and the previous linear mapping is that here there is a bias towards lower values of the entropy, that is, higher entropies are penalized more. But what would be the right way to map entropies to weights?

3.4 Flexible weighting based on entropy

The non-linear mapping presented earlier has a drawback in the fact that, for large entropy values, a large variation in entropy translates into a small variation for the weights. The mapping is less sensitive to variations in entropy where the entropy is higher. Intuitively, the mapping should be more sensitive for some entropy value intervals compared to others, and those "sensitive" intervals should be the ones that include the entropy values that occur most often. This intuition lead to the following method of selecting the mapping.

First, a histogram of past entropy values is built for both streams. In our case, the histogram has 15 bins and comprises 150 past entropy values from both streams, for a total of 300 samples. Then, a piecewise-linear function is built, mapping low entropy values to high weights and vice-versa. This is done in such a way that the slope of each piece is proportional to the number of points contained in the corresponding histogram bin. Figure 1 shows an example of such a mapping and the histogram from which it was built.

This mapping is itself dynamic. It adapts to the particular configuration of entropy values, with the purpose of having the best discrimination power between the most occurring ones.

4. IMPLEMENTATION DETAILS

For our experiments, we use sequences from the CUAVE audio-visual database [15]. They consist of 36 speakers repeating the 10 digits. We use only the static part of the database, that is, the first 5 repetitions.

The video sequences are filmed at 30 fps interlaced, so we can effectively double this framerate through deinterlacing. The average length of one video sequence is around 50 seconds (3000 deinterlaced frames).

Out of the 36 sequences, 30 are used for training, and 6 for testing. We use a six-fold crossvalidation procedure, that is, we repeat training and testing 6 times, each time changing the respective sets using a circular permutation. The performance reported is the average on the 6 runs.

We start our visual feature processing by locating the region of the mouth, scaling and rotating it, such that all the mouths have more or less the same size and position. The temporal resolution of the video is then increased through interpolation, to reach 100 fps, since synchrony between the audio and the video streams is required by our integration method.

The visual features that we use are even-frequency discrete cosine transform (DCT) coefficients of the mouth images, since they contain the information related to the symmetrical details of the image, as detailed in [16]. From them, the highest-energy 64 coefficients are selected, with their

first and second temporal derivatives, and LDA is applied on them, to obtain a 40-dimensional feature vector.

On the audio side, the features extracted are 13 Mel Frequency Cepstral Coefficients (MFCCs), together with their first and second temporal derivatives. Audio features are extracted 100 times per second, at the same frequency as the visual features. Different levels of white gaussian noise are added in order to show how our dynamic weighting algorithm performs across a large range of SNRs.

We use the HTK library [11] for the HMM implementation. Our word models have 8 states with one diagonal-covariance gaussian per state. The silence model has 3 states with 3 gaussians per state. Two streams are used, audio and video. The grammar consists of any combination of digits with silence in-between. The accuracy that we report is the number of correctly recognized words minus insertions, divided by the total number of test words.

There are two possible ways to train multi-stream HMMs. The first one is separate training, where different models are built and trained for each modality. The two resulting HMMs are then merged into a multi-stream HMM, containing the gaussian mixtures from both original models. However, there is no guarantee that the models will be trained on the same alignment of audio and video, so the states might be poorly synchronized.

The second method of training ensures that each HMM state will be trained on the same segment of speech in both modalities. This can be achieved by using a joint multi-stream model from the beginning. However, the problem that arises here is the choice of the weights used in training. We decided to use this second method in our experiments, with both weights equal to 0.5 for training. We found that the initial choice of weight has little influence on the result, and that, for low SNRs, the jointly trained models perform better than the separately trained ones.

5. EXPERIMENTAL RESULTS AND DISCUSSION

In figure 2, we present our results with static weights, that is, the optimal weights for a certain SNR. These optimal weights were found by running the recognizer with a range of possible weight values and then choosing the one that gives the best performance. However, this contradicts our initial assumption that the noise level is unknown at the time of training, so these performance values are only given as an indicator of what the performance could be, ideally.

Our results show that with optimally picked weights, the audio-visual speech recognizer always outperforms both single modality ones, and by a large margin. The visual-only recognition rate is only 54.8%, and still, across all SNRs, there is a significant gain from putting together the two modalities. For example, while the audio-only recognition rate at -10dB is only 38%, the audio-visual performance is much higher, at 67.2%.

We performed tests with several dynamic weighting strategies. We started with the negative entropy weighting described in section 3.2 and the inverse entropy one from section 3.3. We then tested our flexible entropy to weight mapping presented in section 3.4. We also performed tests with another method from the literature, the maximum stream posterior (MSP) method presented in [6], for comparison purposes. We show our results in table 1 and figure 3.

Our results show more or less the same trends with all

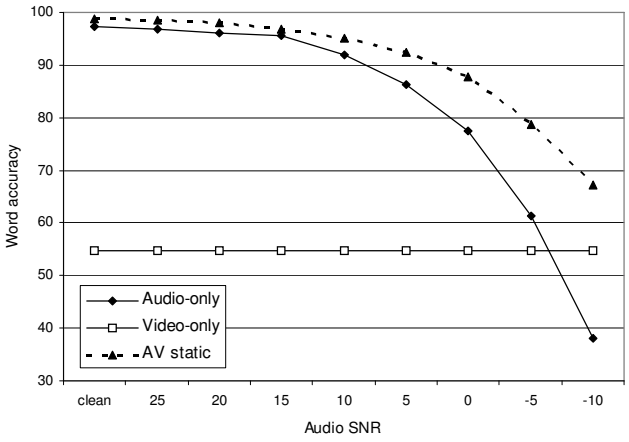

Figure 2: Audio-visual speech recognition performance with static weights, compared to audio-only and video-only results.

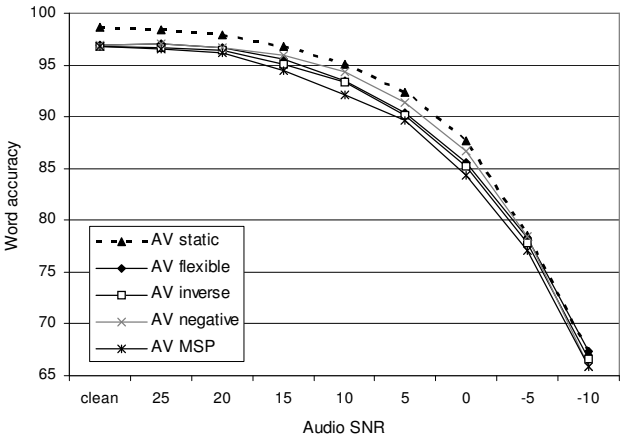

Figure 3: Audio-visual speech recognition performance with static and dynamic weights.

SNR(dB)	Audio-only	static	flexible	inverse	negative	MSP
clean	97.31	98.66	96.93	96.82	96.88	96.76
25	96.91	98.44	96.99	96.65	97.10	96.54
20	96.12	97.93	96.71	96.48	96.65	96.15
15	95.62	96.81	95.53	95.02	95.93	94.47
10	91.92	95.02	93.46	93.35	94.30	92.12
5	86.22	92.34	90.37	90.15	91.39	89.61
0	77.54	87.65	85.53	85.20	86.71	84.41
-5	61.36	78.60	78.31	77.81	78.39	77.11
-10	38.04	67.17	67.33	66.60	65.87	65.86

Table 1: Speech recognition performance for audio-only recognition and several audio-visual integration methods.

dynamic weighting strategies. First, the performance of the dynamic weighting methods is quite close to that of the fixed weighting strategy, with a maximum difference of around 2%. The worst loss of performance is incurred when the audio is clean. This could be explained by the fact that the dynamic weights mean is never biased strongly enough in favor of one modality. In the case of clean audio, the ideal fixed weight is 0.9 for the audio, while the mean of the dynamic weight is closer to 0.75. However, for lower SNRs, dynamic weighting methods are edging closer to the performance of the fixed weights, with the negative entropy method perform-

SNR(dB)	static	flexible	negative	inverse	MSP
15	96.48	94.63	94.97	94.13	93.97
0	80.23	80.50	81.17	79.89	79.16

Table 2: Audio-visual speech recognition performance for static and dynamic weighting methods, with time-varying noise.

ing best. This trend continues all the way to -10dB, where, nevertheless, the flexible weighting method outperforms all the others. The constantly under-performing method is the MSP, but, in the end, the differences are quite small.

The final experiment that we performed is with time-varying noise. Here, the power of the noise signal was changed randomly every one second. We imposed two SNR levels on the final signal, 15dB and 0dB. The interval of variation of the SNR is around 10dB, above and below the mean. This is intended as a more realistic testing scenario, since in real life the SNR is not expected to remain constant.

The results are presented in table 2. As can be seen the same trends as in the constant noise case are present. MSP still performs worse than any of the entropy-based methods. The negative entropy is the best-performing at 0dB, surpassing even the static weights method. It should be mentioned that with time-varying noise the error rate is much higher than with constant noise. This is because the error rate does not vary linearly with the SNR, that is, for the time intervals where the SNR is really low, a lot more is lost with respect to the average accuracy than it is gained where the SNR is high.

The time-varying noise results prove that using a dynamic weighting scheme can lead to good performance even when the SNR changes unexpectedly. The weights are able to adapt quickly to changes in the noise level, adjusting the relative importance of each modality automatically.

6. CONCLUSIONS AND FUTURE WORK

We have presented several methods of audio-visual integration at the early stage of the recognition process, with HMMs synchronized at state level. Our method of using the entropy as a reliability estimate leads to good results across different noise conditions, and better than those obtained with other methods from the literature.

As future work, the application of state-dependent weights might improve the recognition rates. It is well known that some sounds are highly confusable in either the audio or the video modalities. Adjusting the weights in such a way that the modality with higher discriminative power is favored for a certain HMM state should increase the overall performance.

Acknowledgement

This work is supported by the Swiss National Science Foundation through the IM2 NCCR.

REFERENCES

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in audio-visual speech processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds. MIT Press, 2004.

[2] G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, "Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR," *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2002.

[3] S. Cox, I. Matthews, and A. Bangham, "Combining noise compensation with visual information in speech recognition," *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, 1997.

[4] A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by humans and machines*, D. G. Stork and M. E. Hennecke, Eds. Springer, 1996, pp. 461–471.

[5] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000.

[6] R. Seymour, D. Stewart, and J. Ming, "Audio-visual integration for robust speech recognition using maximum weighted stream posteriors," *Proc. INTERSPEECH*, 2007.

[7] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77(2), 1989.

[8] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proceedings of the International Conference on Image Processing*, vol. 3, 1998, pp. 173–177.

[9] R. Reilly and P. Scanlon, "Feature analysis for automatic speechreading," *Proc. Workshop on Multimedia Signal Processing*, pp. 625–630, 2001.

[10] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. International Conference on Spoken Language Processing*, pp. 426–429, 1996.

[11] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, Entropic Ltd., 1999.

[12] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[13] K. Kirchhoff and J. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," *Proceedings ICASSP-99*, pp. 693–696, 1999. [Online]. Available: citeseer.ist.psu.edu/kirchhoff99dynamic.html

[14] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2003.

[15] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP JASP*, vol. 2002(11), pp. 1189–1201, 2002.

[16] P. Scanlon and G. Potamianos, "Exploiting lower face symmetry in appearance-based automatic speechreading," *Proc. Works. Audio-Visual Speech Process. (AVSP)*, pp. 79–84, 2005.