INTERSPEECH 2009 BRIGHTON

Human Audio-Visual Consonant Recognition Analyzed with Three Bimodal Integration Models

Zhanyu Ma and Arne Leijon

Sound and Image Processing Lab,
KTH - Royal Institute of Technology
Stockholm, Sweden
{zhanyu.ma, arne.leijon}@ee.kth.se

Abstract

With A-V recordings, ten normal hearing people took recognition tests at different signal-to-noise ratios (SNR). The A-V recognition results are predicted by the fuzzy logical model of perception (FLMP) and the post-labelling integration model (POSTL). We also applied hidden Markov models (HMMs) and multi-stream HMMs (MSHMMs) for the recognition. As expected, all the models agree qualitatively with the results that the benefit gained from the visual signal is larger at lower acoustic SNRs. However, the FLMP severely overestimates the A-V integration result, while the POSTL model underestimates it. Our automatic speech recognizers integrated the audio and visual stream efficiently. The visual automatic speech recognizer could be adjusted to correspond to human visual performance. The MSHMMs combine the audio and visual streams efficiently, but the audio automatic speech recognizer must be further improved to allow precise quantitative comparisons with human audio-visual performance.

Index Terms: Audio-visual recognition, Fuzzy Logical Model of Perception, Post-Labelling Model, Hidden Markov Models, Multi-Stream Hidden Markov Models

10.21437/Interspeech.2009-60

1. Introduction

Although human beings are good at speech recognition, it is still very difficult to understand the spoken contents in a noisy background. Several experiments have shown that visual motion of lips can help the listener to understand. In a noisy environment, the benefit from the visual signal is larger at a low acoustic signal-to-noise ratio (SNR). How well do human observers combine the audio cues and visual cues while listening in a noisy background and how much benefit can they gain from this combination?

The main goal of this work is to evaluate computational models of bimodal speech integration. As the computational models assume ideal bimodal integration, a comparison between the human beings and the optimal models can indicate the efficiency of the human bimodal integration.

Some well-known models for audio-visual integration were introduced in the past decades. Massaro [1] elaborated principles and methods for audio-visual integration in general. Braida [2] and Grant [3] analyzed predicted and observed integration accuracy with the fuzzy logical model of perception (FLMP) [4], the post-labelling integration model (POSTL) and the pre-labelling integration model (PRE). These three models use different methods to predict audio-visual consonant recognition, using experimental consonant confusion matrices obtained with human observers in audio-only and visual-only test conditions.

Thanks to Martin Dahlquist for supplying the speech test material.

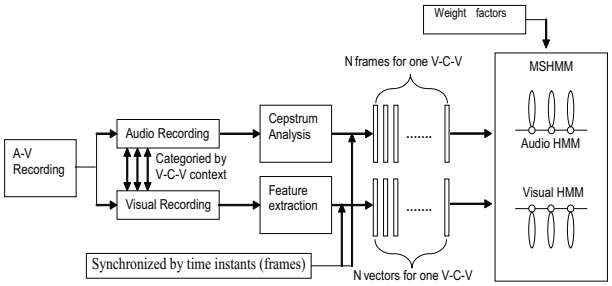


Figure 1: *MSHMM architecture.*

The FLMP and POSTL use the discrete confusion-matrix data directly. The PRE model assumes that the speech-recognition performance is determined by the probability distributions of some hypothetical multidimensional speech features, but the feature distributions are estimated from the confusion matrices by multidimensional scaling (MDS), and have no connection with the physical signal features.

As an alternative to the PRE model, we use the real physical signal features and apply automatic speech recognition methods for comparison with human unimodal and bimodal performances. For this purpose, we train two sets of unimodal hidden Markov models (HMMs) and a set of multi-stream hidden Markov models (MSHMMs) [5, 6] to simulate the process of how humans integrate the audio and visual streams while listening.

Also, using recorded nonsense vowel-consonant-vowel ($V-C-V$) words, we measure the consonant recognition of normal hearing persons in one visual-only condition, and in audio-only and audio-visual conditions with different levels of noise. The results will be compared with the prediction obtained by the models mentioned above.

2. Three Bimodal Integration Models

2.1. Multi-Stream Hidden Markov Model

As an alternative to the pre-labelling model, we applied HMMs to model the dynamic properties of the audio and visual signal features. Also, multi-stream hidden Markov models (MSHMMs) were utilized to simulate the process of audio-visual in-

tegration. The HMM approach has been very successful for automatic speech recognition. The MSHMM approach is widely used in audio-visual speech recognition. With the assumption that, at a given state, the audio signal is conditionally independent from the visual signal to the same stimulus, we simply multiply the output likelihoods of the audio and visual observations together as

$$b_{AV,j}(O_{a,t}, O_{v,t}) = b_{A,j}^{\lambda_A}(O_{a,t}) \cdot b_{V,j}^{\lambda_V}(O_{v,t}) \quad (1)$$

In (1), $b_{A,j}(O_{a,t})$ and $b_{V,j}(O_{v,t})$ are the likelihood of observed feature vectors $O_{i,t}, i \in (A, V)$ given state $S_t = j$ at time instant t, and $\lambda_i, i \in (A, V)$ is the stream weighting factor for the ith stream. At each HMM state in both the A and V models, a Gaussian mixture model (GMM) is used as the observation probability density function (PDF). As described in [5, 6], λ_i might be arbitrary chosen (e.g. $0 \leq \lambda_i \leq 1$, $\sum_{i=a,v} \lambda_i = 1$). In our case, to simulate optimal integration, we assume that both the stream weights were equal to 1.

In our analysis, we have three sets of HMMs, one for each nonsense word. One set is for audio training and recognition, the second one is for the visual stream and the third one is a set of MSHMMs using both the audio and visual streams.

Twelve mel-frequency cepstral coefficients (MFCCs) were used as audio features. For each video frame, we cropped the speaker's mouth area and rearranged the pixel values into a vector. These vectors were used as visual feature vectors. As in [7], probabilistic principal component analysis (PPCA) based GMMs [8, 9] were used for the visual HMMs and the visual part of MSHMMs. For the visual part, the PPCA automatically decreases the underlying dimensionality of the feature space and keeps only the principal components. Fig. 1 shows the architecture of the MSHMMs.

2.2. Fuzzy Logical Model of Perception

In the FLMP [4], the audio-visual (A-V) response to each stimulus is determined in a probabilistic way. For each unimodal stream, the conditional probability of response R_n given the stimulus S_m is denoted as $P_i(R_n|S_m), i \in (a, v)$. The response probabilities for the bimodal stimuli are calculated as

$$P_{a,v}(R_n|S_m) = \frac{P_a(R_n|S_m)P_v(R_n|S_m)}{\sum_{k=1}^{|\mathcal{A}|} P_a(R_k|S_m)P_v(R_k|S_m)} \quad (2)$$

We estimate the confusion probabilities by the MAP method as

$$P_i(R_n|S_m) = \frac{h_n + 1}{H + |\mathcal{A}|} \quad (3)$$

where h_n means the number of choices to response n given stimulus S_m, H denotes the total number of responses given stimulus S_m, and $|\mathcal{A}|$ is the total number of response alternatives.

2.3. Post-labelling Integration Model

According to the post-labelling integration model (POSTL) [2], the individual is assumed to make independent recognition with the audio stream and visual stream. Then the listener combines these two hard judgements to decide the final response to the bimodal stimulus. Each stimulus S_m is assumed to generate a pair of discrete labels (A_p, V_q) with probability mass $P_{a,v}^*(A_p, V_q|S_m)$. The set $\mathcal{U}_n$ of label pairs (p, q) that lead to the same response R_n is used to calculate the probability of the response R_n given the stimulus S_m

$$P_{a,v}(R_n|S_m) = \sum_{(p,q)\in\mathcal{U}_n} P_{a,v}^*(A_p, V_q|S_m) \quad (4)$$

The POSTL assumes that the labels from one stream only depend on this stream and are independent of the other stream. Thus the label probability for each stream can be multiplied to give the multi-stream probabilities as

$$P_{a,v}^*(A_p, V_q|S_m) = P_a^*(A_p|S_m) \times P_v^*(V_q|S_m) \quad (5)$$

The probability of the labels in each stream is estimated from the observed confusion matrices as

$$P_a^*(A_p|S_m) = P_a(R_p|S_m) \quad (6)$$

and similarly for the visual stream, using (3). Among many possible rules that the listener could use to map the label pair (A_p, V_q) to a response, the maximum likelihood rule gives the greatest recognition score when the prior probabilities of the stimuli are equal. For each (A_p, V_q), the response is the identity of the stimulus for which $P_{a,v}^*(A_p, V_q|S_m)$ is greatest. Thus, $\mathcal{U}_n$ is chosen as the set of label pairs (p, q) for which $P_{a,v}^*(A_p, V_q|S_n) > P_{a,v}^*(A_p, V_q|S_k)$ for all $k \neq n$. The conditional response probabilities under the maximum likelihood rule are obtained as

$$P_{a,v}(R_n|S_m) = \sum_{(p,q)\in\mathcal{U}_n} P_a(R_p|S_m) \times P_v(R_q|S_m) \quad (7)$$

If there is a label pair (p, q) for which

$$P_{a,v}^*(A_p, V_q|S_l) = P_{a,v}^*(A_p, V_q|S_n) > P_{a,v}^*(A_p, V_q|S_k) \quad (8)$$

for all $k \neq l, n$, this pair could be arbitrarily assigned to R_l or R_n.

3. Experiments

3.1. Material and Procedures

The consonant test material was recorded in Swedish and contains 17 consonants, /m, b, p, v, f, n, d, t, s, j, ʃ, ç, r, l, g, k, ŋ/ represented in /a/−C−/a/, /i/−C−/i/, and /u/−C−/u/ context. The test material was video-recorded, with the speaker in a frontal position. The Long Time Average Speech Spectrum (LTASS) was measured for all recorded speech, and stationary noise with this spectrum was added to the acoustic channel of the recording while keeping the visual channel undistorted and synchronized. Noise was added to the clean speech at signal to noise ratios (SNR) of 0 dB, -3 dB, -6 dB and -9 dB. Each subject was tested in audio-only and audio-visual conditions at each SNR, and once in the visual-only condition. At a specific SNR, for one vowel, for example /a/, all the /a/−C−/a/ nonsense words were played 3 times in a random order. Thus, in the audio-only condition, $17 \times 3 = 51$ /a/−C−/a/ nonsense words were presented to the subject. The subject was required to respond by choosing one consonant (*e.g.* /m/) immediately after the word was played. All the response alternatives were presented on the computer screen, placed in a fixed layout. Ten people with normal hearing and vision volunteered to participate in the test.

3.2. Test Result Analysis

Test results were collected in the form of confusion matrices. In addition to overall recognition scores we use the Mutual Information (MI) to describe the consonant confusion matrices. The MI is a measure of the mutual dependence between two variables. In [10] it is used to measure the dependence between the stimuli and responses in consonant recognition tests. In our experiment, the stimulus is a discrete variable S with 17 possible values with equal probability. Also, the response R to the stimulus is another discrete variable with 17 possible choices. In the confusion matrix, each cell counts the number of responses when a specific stimulus is given. Using (3), the response frequencies were used to estimate the probabilities $P(R_n|S_m)$.The MI between R and S is then

$$I(R;S) = \sum_{s\in S} P(s) \sum_{r\in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} . \qquad (9)$$

4. Results and Discussion

All the ten participants finished all the tests. Using the observed audio-only and visual-only confusion matrices, we predicted the A-V result via the FLMP and the POSTL models.

For HMMs and MSHMMs, we trained and tested them with the audio materials, visual materials and the audio-visual materials. Separate models were trained for each nonsense word, including all the three sets of nonsense consonant words at all the SNR levels. When testing at a specific SNR, we selected one set of data from the training data, obtained one set of result, and then chose the next set till all the three sets were tested. The final results are the mean of these three.

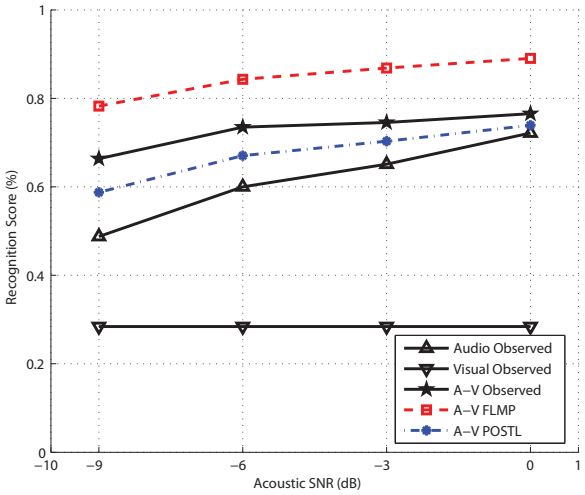
4.1. Recognition Scores and Mutual Information

The observed performance by human participants and by the FLMP and POSTL models are shown in Fig. 2. As the noise level increases, the audio recognition score of the participants becomes worse, as expected. As shown in Fig. 3, the audio-only HMMs perform worse than human beings at each SNR. The visual-only HMMs perform better than the human participants. To facilitate comparison, we artificially increased the variance of the output PDFs in the visual-only HMMs to reduce the recognition score of the visual-only HMMs to the same level as the human performance. The same modification was made in the visual part in the MSHMMs, which decreases the A-V recognition score. Fig. 3 shows the recognition scores of the MSHMMs both before and after this modification.

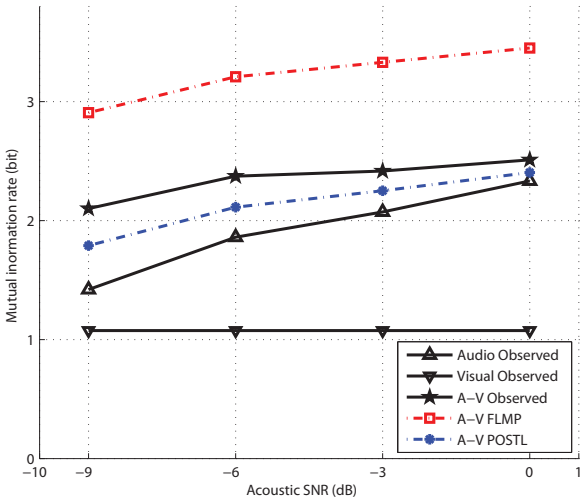
4.2. Performance Analysis of the Models

The observed and estimated recognition scores are listed in table 1. Fig. 2 shows that both the audio-visual integration models (FLMP and POSTL) predict the main trends of how the recognition scores change under different SNRs. However, the FLMP predicted much better A-V results than observed, and the POSTL underestimates human A-V performance.

With the FLMP, the mutual information between stimulus and response in the A-V condition was even slightly greater than the sum of the mutual information in the unimodal conditions (Fig. 2(b)). This is clearly theoretically incorrect, as the audio and visual channels convey partly redundant speech information, which is also confirmed by the observed human results. For the POSTL, the hard decisions for each stream under unimodal conditions are combined to yield the final bimodal response. This is not the optimal integration approach because some information is lost in the hard decisions.



(a) *Recognition score comparison.*



(b) *Mutual information comparison.*

Figure 2: *Observed and model-predicted recognition scores (a) and mutual information between stimuli and responses (b), as a function of SNR, using FLMP and POSTL to predict audio-visual (A-V) performance.*

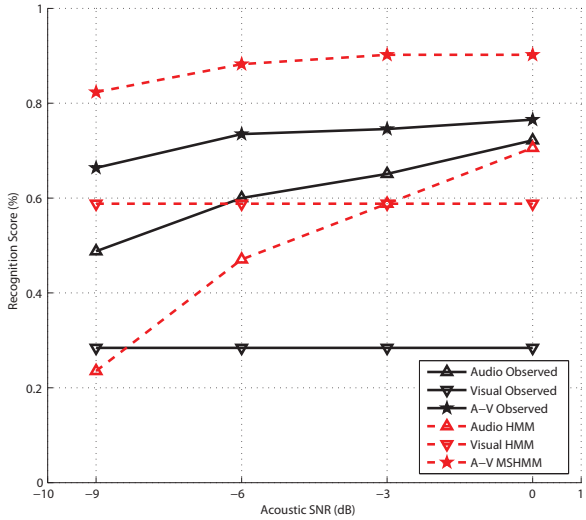
The visual stream is not distorted, while the audio stream is affected by different levels of noise. In the process of integration, MSHMMs make a balance between the audio and visual streams to achieve a better bimodal result. The information from the visual stream compensates the lack of reliability in the noisy audio stream. The overestimation here is due to the robust recognition ability in the visual HMMs. Acoustic noise still affects the final result when the audio stream becomes more unreliable as the SNR decreases.

The modified visual-only HMMs have the same recognition score as the human observers. In this case, with the audio part unchanged, the MSHMMs still integrate the audio stream and visual stream efficiently, but the model underestimates human A-V performance due to the worse performance of the audio-only HMMs (and audio part in MSHMMs). In future work, we will attempt to improve the performance of the audio-only HMMs, and then the integrated result could be more directly comparable to human results.

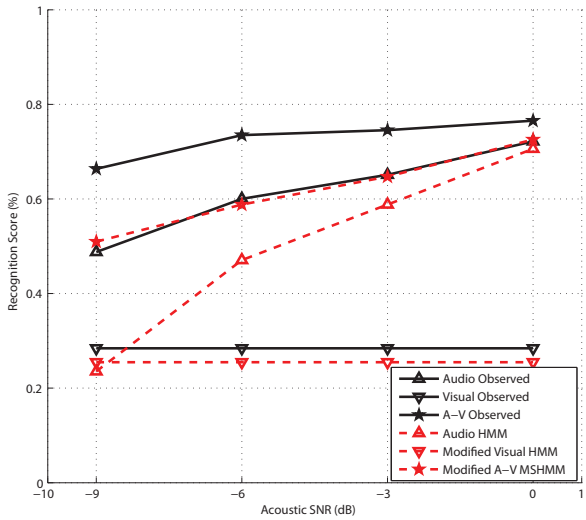
Table 1: *Observed and predicted recognition scores (%)*

	Audio				Visual	Audio-Visual			
	0dB	−3dB	−6dB	−9dB		0dB	−3dB	−6dB	−9dB
Observed	72	65	60	49	28	77	75	74	66
FLMP	N/A*	N/A	N/A	N/A	N/A	89	87	84	78
POSTL	N/A	N/A	N/A	N/A	N/A	74	70	67	59
MSHMM/HMM	72	65	60	49	59	90	90	88	82
Modified MSHMM/HMM**	72	65	60	49	25	73	65	59	51

**For the perception models, there is no prediction for the audio or visual recognition score.*
***Only modified the visual HMM and the MSHMM.*



(a) Observed and Original MSHMM/HMM recognition scores.



(b) Observed and Modified MSHMM/HMM recognition scores.

Figure 3: *Observed and machine estimated recognition scores, as a function of SNR, using trained MSHMMs/HMMs and artificially modified MSHMMs/HMMs to reduce their visual performance.*

5. Conclusion

The visual motions of lips indeed help people and machines to understand the spoken content while listening in noisy environments. We analyzed human audio-visual consonant recognition using three well-known models: FLMP, POSTL, and automatic speech recognition with HMMs and MSHMMs. All the models agree qualitatively with the result that the benefit gained from the visual signal is larger at lower acoustic SNRs. However, the FLMP severely overestimates human audio-visual performance, while the POSTL model underestimates it. With the FLMP, the mutual information between stimulus and response in the A-V condition is close to the sum of the audio-only and visual-only mutual information, which is theoretically incorrect.

The visual automatic speech recognizer could be adjusted to correspond to human visual performance. The audio-visual MSHMMs combine the audio and visual streams efficiently, but the audio automatic speech recognizer must be further improved to allow precise quantitative comparisons with human audio-visual performance.

6. References

[1] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech," *American Scientist*, vol. 86 n3, p. 236, 1998.

[2] L. D. Braida, "Crossmodal integration in the identification of consonant segments," *The Quarterly Journal of Experimental Psychology*, vol. 43A(3), pp. 647–677, 1991.

[3] K. W. Grant, B. E. Walden, and P. F. Seitz, "Auditory-visual speech recognition by hearing-impaired subjuects: Consonant recognition, sentence recognition, and auditory-visual integration," *Journal of Acoustical Society of America*, vol. 103, pp. 2677–2690, 1998.

[4] D. W. Massaro, *Speech perception by ear and eye: A paradigm for psychological inquiry*. Lawrence Erlbaum Associates, Inc, 1987.

[5] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 2021 – 2024, 2002.

[6] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 17–21, 2004.

[7] Z. Ma and A. Leijon, "A probabilistic principal component analysis based hidden markov model for audio-visual speech recognition," in *Proceedings of Asilomar Conference on Signals, Systems & Computers*, 2008.

[8] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11(2), pp. 443–482, 1999.

[9] ——, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society*, vol. 21(3), pp. 611–622, 1999.

[10] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *Journal of the Acoustical Society of America*, vol. 27, p. 338, 1955.