論文 / 著書情報
Article / Book Information

Title	A Robust Multi-Modal Speech Recognition Method Using Optical-Flow Analysis
Authors	Satoshi Tamura, Koji Iwano, Sadaoki Furui
Citation	ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments, Vol. , No. , pp. 2-4
Pub. date	2002, 6

Proc. ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments, Kloster Irsee, Germany (2002)

A ROBUST MULTI-MODAL SPEECH RECOGNITION METHOD USING OPTICAL-FLOW ANALYSIS

Satoshi Tamura, Koji Iwano and Sadaoki Furui
Department of Computer Science, Tokyo Institute of Technology
{tamura,iwano,furui}@furui.cs.titech.ac.jp

Abstract This paper proposes a new multi-modal speech recognition method using optical-flow analysis, evaluating its robustness to acoustic and visual noises. Optical flow is defined as the distribution of apparent velocities in the movement of brightness patterns in an image. Since the optical flow is computed without extracting speaker's lip contours and location, robust visual features can be obtained for lip movements. Our method calculates a visual feature set in each frame consisting of maximum and minimum values of integral of the optical flow. This feature set has not only silence information but also open/close status of the speaker's mouth. The visual feature set is combined with an acoustic feature set in the framework of HMM-based recognition. Triphone HMMs are trained using the combined parameter set extracted from clean speech data. Two multi-modal speech recognition experiments have been carried out. First, acoustic white noise was added to speech wave forms, and a speech recognition experiment was conducted using audio-visual data from 11 male speakers uttering connected Japanese digits. The following improvements of relative reduction of digit error rate over the audio-only recognition scheme were achieved, when the visual information was incorporated into silence HMM: 32% at SNR=10dB and 47% at SNR=15dB. Second, a real-world data distorted both acoustically and visually was recorded in a driving car from six male speakers and recognized. We achieved approximately 17% and 11% relative error reduction compared with audio-only results on batch and incremental MLLR-based adaptation, respectively.

Keywords: multi-modal speech recognition, optical flow, robust to noise, speaker independent

1. INTRODUCTION

Automatic Speech Recognition (ASR) systems are expected to play important roles in an advanced multi-media society with user-friendly human-machine interfaces such as ubiquitous computing environments

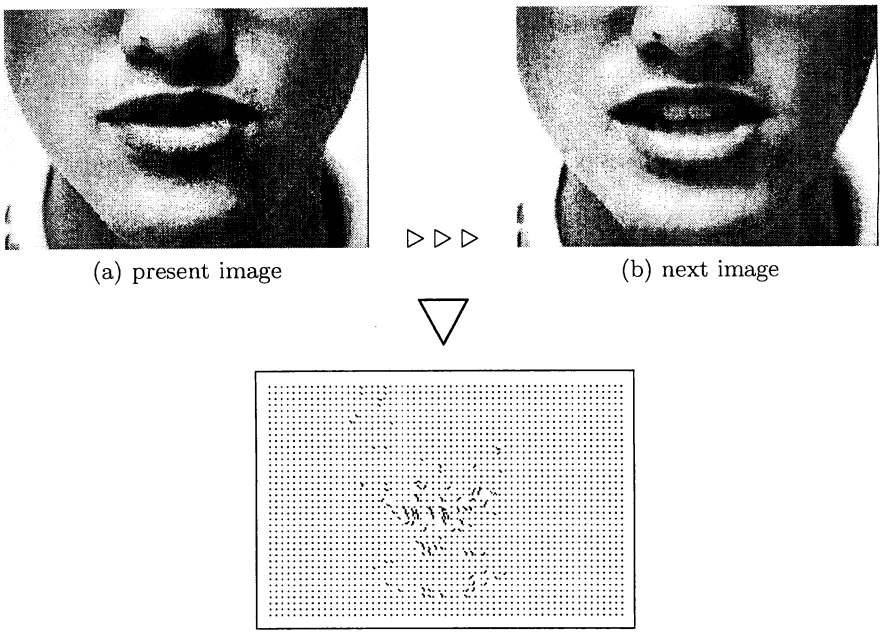
[Furui et al., 2001]. High recognition accuracy can be obtained for clean speech using the state-of-the-art technology even if the vocabulary size is large, however, the accuracy largely decreases in noisy conditions. Therefore, increasing the robustness to noisy environments is one of the most important issues of ASR.

Multi-modal speech recognition, in which acoustic features and other information are used jointly, has been investigated and found to increase robustness and thus improve the accuracy of ASR. Most of the multi-modal speech recognition methods use visual features, typically lip information, in addition to the acoustic features [Nakamura et al., 2000; Miyajima et al., 2000; Potamianos et al., 1997; Bregler and Konig, 1994]. By using the visual information, acoustically similar sounds, such as nasal sounds: /n/, /m/, and /ng/, become easier to recognize [Basu et al., 1999; Mase and Pentland, 1991]. In most of the studies, a lip is found from an image by mouth tracking, subsequently the lip contour is extracted, and visual features are obtained by pattern matching techniques or signal processing methods such as FFT. Since it is not easy to determine a mouth location and extract a lip shape, lip marking is often needed to ensure robust extraction of visual features in these schemes.

Mase and Pentland reported their lip-reading system for recognizing connected English digits using an optical-flow analysis [Mase and Pentland, 1991]. Optical flow is defined as the distribution of apparent velocities in the movement of brightness patterns in an image [Horn and Schunck, 1981]. The following advantages exist with using the optical flow for audio-visual multi-modal speech recognition. First, the visual features can be detected robustly without extracting lip locations and contours. Second, it is more reasonable to use lip motion for lip reading rather than using a lip shape. Third, the visual features are independent of the speaker's mouth shape or beard.

We have proposed a multi-modal speech recognition scheme using the optical-flow analysis for extracting visual information [Iwano et al., 2001]. We have used variances of horizontal and vertical elements of optical-flow vectors as a visual feature set, and found that they are especially useful for estimating pause/silence periods. We achieved about 30% relative error reduction compared with audio-only results when recognizing white-noise-added speech at 10dB SNR level condition. However, robustness of the proposed method to visual noise using an audio-visual data in real environments has not yet been evaluated. Increasing the visual robustness is crucial to make the method applicable to mobile environments.

In this paper, we conduct recognition experiments for not only artificially noise-added speech but also real-world speech using a new visual



(a) present image (b) next image

(c) optical-flow vectors

Figure 1. An example of optical-flow analysis

feature extraction method. We describe recognition results comparing with results with the audio-only recognition scheme, and evaluate both acoustic and visual robustness of our method. This paper is organized as follows: In Section 2 the principle of the optical-flow method is explained. Our audio-visual multi-modal speech recognition system is described in Section 3. Experimental setup and results for acoustic noise-added data are shown in Section 4, and for real-world data are in Section 5. Finally we conclude our research and describe our future works in Section 6.

2. OPTICAL-FLOW ANALYSIS

Optical flow is the distribution of apparent velocities in the movement of brightness patterns in an image. We use the Horn-Schunck algorithm [Horn and Schunck, 1981]. This algorithm has an advantage that it needs no characteristic point in contrast with pattern-matching-based algorithms, and that it requires only two images in processing. In this method, brightness at every point is assumed to be constant during movement for a short time. From this assumption, the time derivative

of the brightness is zero:

$$\frac{dI}{dt} \simeq \frac{\partial I}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial I}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \tag{1}$$

where $I(x, y, t)$ is a brightness of a point (x, y) in an image at time t. In this equation (1), if we let

$$u = \frac{dx}{dt} \quad \text{and} \quad v = \frac{dy}{dt} \tag{2}$$

then the following constraint is obtained:

$$I_x \cdot u + I_y \cdot v + I_t = 0 \tag{3}$$

Here $u(x, y)$ denotes a horizontal element of optical flow at a point (x, y), and $v(x, y)$ denotes a vertical element. Since we cannot determine $u(x, y)$ and $v(x, y)$ using only the equation (3), we incorporate another restraint which minimizes the sum of the square values of the magnitude of the gradient of $u(x, y)$ and $v(x, y)$ at every point:

$$\iint \{({u_x}^2 + {u_y}^2) + ({v_x}^2 + {v_y}^2)\}\, dx\, dy \rightarrow \min \tag{4}$$

Then the optical-flow vectors $u(x, y)$ and $v(x, y)$ are computed under these two constraints (3) and (4) by an iterative technique using the average of optical-flow velocities estimated over neighboring pixels. An example of the optical-flow analysis is shown in Figure 1. The left image (a) is extracted from a video sequence at a certain time, and the right image (b) is the next picture. An image of optical-flow velocities computed from these images is shown in (c).

3. A MULTI-MODAL SPEECH RECOGNITION SYSTEM

3.1 Feature extraction and fusion

Figure 2 shows the structure of our audio-visual multi-modal speech recognition system. Speech signals are recorded at a 16kHz sampling rate, and 39-dimensional acoustic features, consisting of 12 Mel-Frequency Cepstral Coefficients (MFCCs), normalized log-energy and their first and second derivatives, are extracted at every 10ms. A video stream is captured with the frame rate of 15 frames/sec and the resolution size of 360×240. Before computing the optical flow, the resolution is reduced to 180×120 keeping the aspect ratio so that computation complexity should be reduced, and the image is transformed into gray-scale. Low-pass filtering (smoothing) and low-level noise addition are applied in

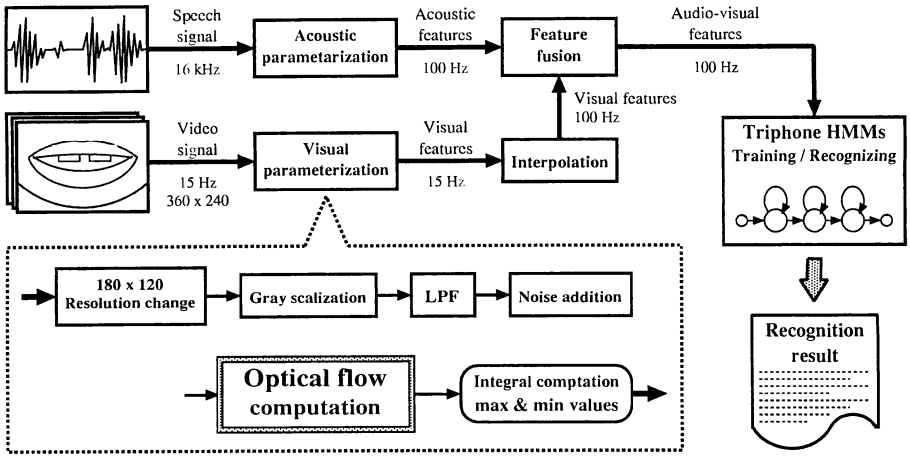
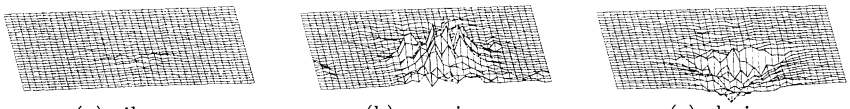


Figure 2. Our multi-modal speech recognition system



(a) silence (b) opening (c) closing

Figure 3. Examples of optical-flow integral results

order to increase the precision of the optical flow. Then the optical flow is computed from a pair of consecutive images with five iterations.

We extract a visual feature set from the optical-flow analysis. It consists of maximum and minimum values of the integral of the optical-flow vectors. The 3-D images of integral results are shown in Figure 3. When a speaker is not speaking, the surface is almost flat as shown in the left image (a). When the speaker's mouth is opening, optical-flow vectors point in diffusing directions around the mouth shape. As a result, a mountain-like surface is created as shown in the center image (b), and it produces the maximum value. When the mouth is closing, converging vectors of optical flow occur around the lip contour. Then the integral operation produces a dip in the mouth area, as shown in the right image (c), and the minimum value is observed. Therefore, this feature set contains not only moving information but also open/close information of the mouth. The 39-dimensional acoustic features and the 2-dimensional visual features are combined into a 41-dimensional audio-visual vector,

after synchronizing the visual frame rate with the audio frame rate using a 3-degree spline function.

3.2 Modeling

A set of triphone Hidden Markov Models (HMMs) having three states and two mixtures in each state is used in our system. After training the audio-visual features with the EM algorithm, the streams of the states in all triphone HMMs are divided into the 39-dimensional audio and 2-dimensional visual streams. The observation probability $b_j(O_{AV})$ of generating an audio-visual feature O_{AV} is given by the following equation:

$$b_j(O_{AV}) = b_{A_j}(O_A)^{\lambda_A} \times b_{V_j}(O_V)^{\lambda_V} \tag{5}$$

where $b_{A_j}(O_A)$ and $b_{V_j}(O_V)$ are probabilities of generating an acoustic vector O_A and a visual vector O_V in a state j respectively, and λ_A and λ_V are weighting factors for the audio and visual streams. By properly controlling these weighting factors according to the noise condition, improvements of the recognition accuracy compared with the audio-only ASR is expected.

4. EXPERIMENTS FOR NOISE-ADDED DATA

4.1 Database

An audio-visual speech database was collected in a clean/quiet condition from 11 male speakers, each uttering 250 sequences of connected digits in Japanese. Each sequence consisted of 2–6 digits, such as "3029 (*san-zero-nī-kyū*)" and "187546 (*ichi-hachi-nana-gō-yon-roku*)", with an average of four digits. The total duration of our database is approximately 2.5 hours.

4.2 Training and recognition

Experiments were conducted using the leave-one-out method: data from one speaker were used for testing while data from other 10 speakers were used for training. This process was rotated for all possible combinations. Since the visual features are considered to be effective especially to detect silence, we controlled the stream weight factors, λ_A and λ_V, only for the silence HMM under the following constraint:

$$\lambda_A + \lambda_V = 1 \quad , \quad \lambda_A \geq 0 \quad , \quad \lambda_V \geq 0 \tag{6}$$

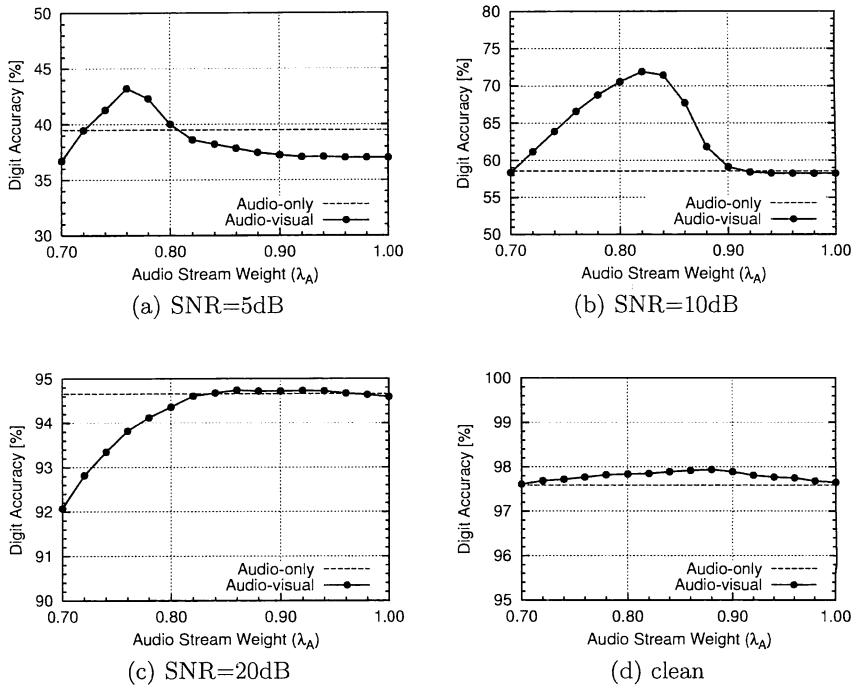


Figure 4. Recognition results for white noise-added speech at various SNR levels

Table 1. The best recognition results for noise-added audio-visual data

SNR	Audio-only	Audio-visual (λ_A)
5dB	39.50%	43.22% (0.78)
10dB	58.55%	71.89% (0.82)
15dB	78.25%	88.21% (0.90)
20dB	94.66%	94.74% (0.86)
clean	97.59%	97.94% (0.88)

For any other triphone HMM, we fixed λ_A and λ_V at 1.0 and 0.0 respectively.

4.3 Results

Figure 4 shows the digit recognition results at (a) 5dB, (b) 10dB, and (c) 20dB SNR level conditions of white noise, and (d) clean condition. The horizontal axis indicates the audio stream weight λ_A, and the vertical axis indicates the percentage of digit recognition accuracy. The

Table 2. The error analysis results (the number of errors) with/without silence evaluations for noise-added audio-visual data consisting of 100 utterances per speaker

		Del	Sub	Ins
Audio-only	(with silence)	14.36	25.01	4.12
	(without silence)	9.64	18.90	12.91
Audio-visual	(with silence)	14.47	9.52	2.35
	(without silence)	14.01	11.83	2.27

(Del: deletion error, Sub: substitution error, Ins: insertion error)

dotted line indicates the accuracy of the audio-only recognition scheme as the baseline, while the solid line indicates the performance of our multi-modal ASR method. Table 1 shows the best multi-modal recognition results and corresponding audio stream weights in comparison with the audio-only results. These results show that our multi-modal ASR system achieves better performance than the audio-only ASR in all environments. Especially, approximately 47% and 32% of relative reduction in the digit error rate, compared with the audio-only recognition scheme, has been achieved in 15dB and 10dB SNR level condition, respectively.

4.4 Considerations

We consider that one of the reasons why the recognition performance was improved is that digit insertion and silence deletion errors were restrained by audio-visual features. In real applications, it is important that digits must not be inserted in pause/silence periods and silences should not be inserted within digit sequences. Since silence deletion, substitution and insertion errors were not counted in the above evaluation, we needed to conduct another evaluation in which silence insertion within speech periods and substitution of silences by digits as well as digit insertion within silences were counted as errors. Table 2 shows the comparison between results of evaluations "with silence" and "without silence" for both audio-only and audio-visual methods at the best stream weight factor $\lambda_A = 0.82$ when SNR=10dB; The table shows the number of errors within 100 utterances per speaker. The difference between the results with/without silence is obvious for substitution and insertion errors in the audio-only condition, whereas there are few differences in the audio-visual condition. This means that silence periods are more correctly detected by the audio-visual method than the audio-only method.

Figure 5. An example of an image in our real-world database (sunlight and car-frame shadow are observed)

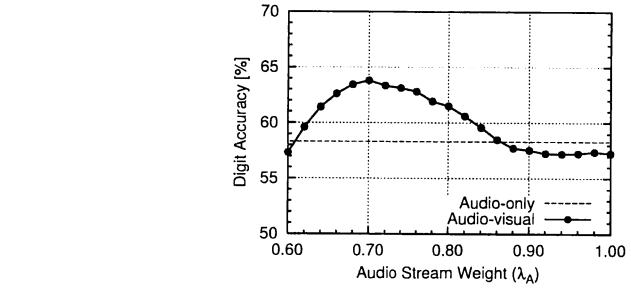
5. EXPERIMENTS FOR REAL-WORLD DATA

5.1 Database

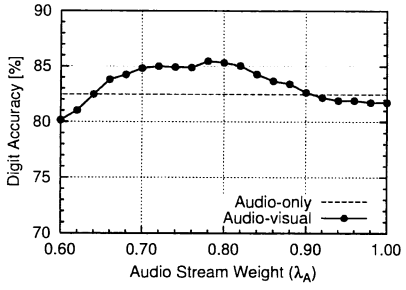
We collected another audio-visual database in a real environment to evaluate both audio and visual robustness of our multi-modal ASR system. Six male speakers different from those in the clean database respectively uttered 115 sequences of connected digits in a driving car on the expressway. The total duration of this database is about an hour. There were several kinds of acoustic noises in our database, such as engine sounds, wind noises, and nonstationary winker sounds. The acoustic SNR level of this database is approximately 10–15dB. As for a visual noise, extreme brightness changing when going through shadows of viaducts and signs, head shaking on a bumpy road, and slow car-frame shadow movement on a face when driving in a curve were observed. An example of visual data in our database is shown in Figure 5.

5.2 Training and recognition

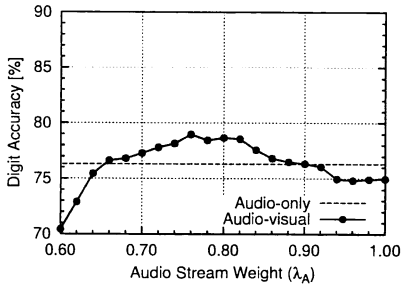
In this experiment, the clean audio-visual database was used for training, while the real-world database was used for testing. The stream weight parameters were restricted by the equation (6). In order to increase the robustness of ASR, Cepstral Mean Normalization (CMN) and unsupervised adaptation using the Maximum Likelihood Linear Regression (MLLR) [Leggetter and Woodland, 1995] technique were applied. The log-energy coefficient was removed from the feature set. The audio-visual feature therefore consisted of 38-dimensional acoustic and 2-dimensional visual features.



(a) No MLLR adaptation



(b) Batch MLLR adaptation



(c) Incremental MLLR adaptation

Figure 6. The recognition results on various MLLR adaptations for real-world data

5.3 Results

Figure 6 shows the recognition results for the real-world audio-visual data recorded in the driving car, under the condition of (a) no adaptation and unsupervised (b) batch/(c) incremental MLLR adaptations: The batch MLLR adaptation is the method in which all test data were used for adaptation before recognition. In the incremental MLLR adaptation, the test set was incrementally used for adaptation; Every time an utterance is recognized, it is then used to update the HMMs. In both methods, not only mean but also variance values of every HMM were adapted, and the number of transformation matrices was three. In the results shown in Figure 6, our system achieved about 17% and 11% relative error reduction compared with the audio-only results on batch and incremental MLLR adaptation, respectively, while about 13% relative error reduction was achieved when no adaptation was applied.

Table 3. The error analysis results (the number of errors) with/without silence evaluations for real-world audio-visual data consisting of 100 utterances per speaker

		Del	Sub	Ins
Audio-only	(with silence)	16.75	8.73	0.94
	(without silence)	2.68	7.83	4.12
Audio-visual	(with silence)	1.46	6.33	4.63
	(without silence)	1.74	7.68	3.19

(Del: deletion error, Sub: substitution error, Ins: insertion error)

5.4 Considerations

We also conducted the evaluation in the same way as in Section 4. Table 3 shows the error analysis results on unsupervised batch MLLR adaptation comparing the audio-only method with our multi-modal ASR system ($\lambda_A = 0.78$). We evaluated only three speakers' speech data since others often uttered isolated digits, inadequate utterances for this analysis. This result is almost as same as the one shown in Table 2. Therefore, it is concluded that the improvement on recognition accuracy is due to the better performance in silence detection by visual features.

6. CONCLUSIONS AND FUTURE WORKS

This paper has proposed a robust visual feature extraction technique for audio-visual multi-modal ASR, and evaluated the robustness of our method against both acoustic and visual noises using real-world data. Our method has achieved the following digit error rate reduction compared with the audio-only schemes: 46% reduction in the white noise condition at 15dB SNR level, and 17% in the real environment on unsupervised batch MLLR adaptation. These experimental results show that our multi-modal ASR system performs well even in noisy conditions such as mobile environments. The visual features are significantly useful for detecting silence and reducing digit insertion errors in silence periods. Since these experiments have been conducted in a speaker-independent condition, it has also been confirmed that our method is effective for speaker-independent tasks.

Our future works include: (1) investigation of more robust and informative visual parameters, such as features including the direction and amount of lip movements, (2) optimization of the stream weight for each triphone HMM to improve the performance by applying the maximum likelihood method or other algorithms, (3) investigation of fusion

algorithm and audio-visual synchronization methods in order to invent robust and high-performance multi-modal ASR techniques, and (4) extension to other tasks or applications such as an information retrieval dialogue system.

ACKNOWLEDGEMENTS

This research has been conducted in cooperation with NTT DoCoMo. The authors wish to express thanks for their support.

References

Basu, S., Neti, C., Rajput, N., Senior, A., Subramaniam, L., and Verma, A. (1999). Audio-visual large vocabulary continuous speech recognition in the broadcast domain. *MMSP'99*, pages 475–481.

Bregler, C. and Konig, Y. (1994). "eigenlips" for robust speech recognition. *ICASSP'94*, 2:669–672.

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., and Tamura, S. (2001). Ubiquitous speech processing. *ICASSP2001*, 1:13–16.

Horn, B. and Schunck, B. (1981). Determining optical flow. *Artificial Intelligence*, 17:185–203.

Iwano, K., Tamura, S., and Furui, S. (2001). Bimodal speech recognition using lip movement measured by optical-flow analysis. *HSC2001*, pages 187–190.

Leggetter, C. and Woodland, P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, pages 171–185.

Mase, K. and Pentland, A. (1991). Automatic lipreading by optical-flow analysis. *Trans. Systems and Computers in Japan*, 22:67–76.

Miyajima, C., Tokuda, K., and Kitamura, T. (2000). Audio-visual speech recognition using mce-based hmms and model-dependent stream weights. *ICSLP2000*, 2:1023–1026.

Nakamura, S., Ito, H., and Shikano, K. (2000). Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. *ICSLP2000*, 3:20–24.

Potamianos, G., Cosatto, E., Gref, H., and Roe, D. (1997). Speaker independent audio-visual database for bimodal asr. *AVSP'97*, pages 65–68.