

Large-Vocabulary Audio-Visual Speech Recognition by Machines and Humans

Gerasimos Potamianos, Chalapathy Neti, Giridharan Iyengar, Eric Helmuth

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598

{gpotam,cneti,giyengar,erichl}@us.ibm.com

Abstract

We compare automatic recognition with human perception of audio-visual speech, in the large-vocabulary, continuous speech recognition (LVCSR) domain. Specifically, we study the benefit of the visual modality for both machines and humans, when combined with audio degraded by speech-babble noise at various signal-to-noise ratios (SNRs). We first consider an automatic speechreading system with a pixel based visual front end that uses feature fusion for bimodal integration, and we compare its performance with an audio-only LVCSR system. We then describe results of human speech perception experiments, where subjects are asked to transcribe audio-only and audio-visual utterances at various SNRs. For both machines and humans, we observe approximately a 6 dB effective SNR gain compared to the audio-only performance at 10 dB, however such gains significantly diverge at other SNRs. Furthermore, automatic audio-visual recognition outperforms human audio-only speech perception at low SNRs.

1. Introduction

During the past decade, solid progress has been made in incorporating visual information extracted from the speaker's lips, or mouth region, into *automatic speech recognition* (ASR) systems, significantly improving their robustness to acoustic degradation [1]-[5]. Research in this area, also known as *automatic speechreading*, although initially limited to single-speaker nonsense, or isolated words [6], has grown to cover multi-speaker connected-word recognition tasks [3], and, recently, speaker-independent, *large-vocabulary, continuous speech recognition* (LVCSR) [5]. For all these tasks, audio-visual ASR has been demonstrated to outperform audio-only ASR, even in the clean audio case. For example, in [5], it is reported that the visual modality reduces the *word error rate* (WER) in LVCSR by 7% relative in clean speech, and by 27% relative in the case where the audio channel is artificially degraded by additive *speech-babble* noise at an 8.5 dB *signal-to-noise ratio* (SNR).

Similarly to automatic speechreading, significant progress has been made in understanding how *humans* perceive audio-visual speech [1], [7]. As a matter of fact, work in human perception has pre-dated and motivated the interest in automatic speechreading. For example, the visual modality benefit to speech intelligibility in noise has been quantified as far back as in 1954 [8], whereas human fusion of audio and visual stimuli has been demonstrated by the *McGurk effect* [9]. There are three reasons why vision benefits human speech perception: Speaker (audio source) localization, segmental information that supplements the audio, and complimentary to the audio speech information, due to the visibility of the place of articulation [10].

Although closely related, joint studies of automatic speechreading and human speech perception are infrequent [11]. Of particular interest to the ASR community, for exam-

ple, would have been to benchmark audio-visual ASR system performance against human listeners, thus possibly identifying areas for its improvement. Although such comparisons have been made for audio-only systems [12], to our knowledge, there exist none for audio-visual word-level recognition [11], [13].

In this paper, and motivated by Lippmann's work [12], we proceed to compare machine and human performance in audio-visual speech recognition. We are particularly interested in the large-vocabulary, continuous speech domain, and in quantifying the visual modality benefit in reducing the WER over audio-only automatic and human recognition at a wide range of acoustic channel SNRs. To achieve this goal, we first consider the automatic speechreading system of [14], and we compare its performance with a traditional audio-only LVCSR system on the IBM ViaVoice™ audio-visual database [5], at a number of SNR levels. Subsequently, we perform a simple human speech perception experiment, where a small number of subjects are asked to transcribe audio-only and audio-visual database utterances at various SNRs. Finally, we compare the WER improvements due to the visual modality, and we report *effective SNR gains* (defined in Section 3) for both machines and humans.

This work is novel for a number of reasons: First, we present new audio-visual ASR results for the automatic speechreading system of [14]: We report recognition performance at a large number of audio SNRs under both matched and unmatched training, obtained by full *decoding*, and by using two feature fusion techniques. In contrast, in [14], only clean and 8.5 dB SNR audio conditions have been considered, and most audio-visual recognition results have been obtained by rescoring audio-only lattices. Second, we present large-vocabulary, continuous audio-visual speech human perception experiments at a number of SNRs, allowing effective SNR gain computations. To our knowledge, human audio-visual LVCSR has also been studied in [15] at a single SNR, whereas effective SNR gains have been reported to be 16 dB for word recognition, in [8], and 11 dB for *white* noise and sentences of limited syntactical variety, in [16]. Last but not least, direct comparisons between machine and human audio-visual LVCSR have never before been performed.

The paper is structured as follows: Section 2 describes the automatic speechreading system adopted. Section 3 presents the audio-visual database and reports audio-only and audio-visual ASR results. Section 4 discusses the human speech perception study, followed by Section 5, that compares machine and human speech recognition. Our conclusions are drawn in Section 6.

2. The automatic speechreading system

A large number of audio-visual ASR systems have been suggested in the literature over the past years [1]-[6]. The two main components that differentiate such systems are [6]: First, the visual front end, namely the extraction of features from the video

of the speaker's face that provide visual speech information, and second, the audio-visual fusion algorithm, i.e., the combination of audio and visual features into a bimodal classifier to recognize audio-visual speech. In this work, we use the automatic speechreading system reported in [14], briefly described next.

2.1. The audio and visual features

We follow the *pixel* (appearance) based approach to visual feature extraction, as opposed to the use of lip-contour (shape) based features, or joint appearance and shape based ones [5], [6]. Given the speaker's video, we first employ a statistical face tracking algorithm to detect the speaker's face and estimate the mouth location and size [5]. Based on these, a size-normalized, 64×64 pixel *region of interest* (ROI) is extracted for every video frame at 60 Hz, containing the speaker's mouth. Subsequently, a two-dimensional, separable, *discrete cosine transform* (DCT) is applied to the ROI, and the 24 highest-energy (over all training data) DCT coefficients are retained as *static* features. To facilitate audio-visual fusion, linear interpolation is used to obtain visual features, time-synchronous to the audio ones at 100 Hz. Finally, *feature mean normalization* (FMN) is employed to compensate for lighting variations, providing the final *visual-only* static features, denoted by $\mathbf{y}_t^{(V)}$. On the other hand, the *audio-only* static features, $\mathbf{y}_t^{(A)}$, consist of 24 mel-frequency cepstral coefficients, computed over a sliding window of 25 msec, at a rate of 100 Hz, after FMN application [14].

To obtain the final audio- and visual-only features, for each modality, a cascade of two linear transforms is applied on a concatenation of consecutive static features, as a means of incorporating *dynamic* speech information and improving recognition. Let the concatenation of J_s consecutive static features be

$$\mathbf{x}_t^{(s)} = [\mathbf{y}_{t-\lfloor J_s/2 \rfloor}^{(s)\top}, \dots, \mathbf{y}_t^{(s)\top}, \dots, \mathbf{y}_{t+\lfloor J_s/2 \rfloor - 1}^{(s)\top}]^\top, \quad (1)$$

with dimension $d_s = J_s n_s$, where $s = A, V$. First, *linear discriminant analysis* (LDA) data projection, and subsequently, a data rotation by means of a *maximum likelihood linear transformation* (MLLT) are applied on vectors $\mathbf{x}_t^{(s)}$. The final audio- and visual-only feature vectors of dimension D_s are denoted by

$$\mathbf{o}_t^{(s)} = \mathbf{P}_{\text{MLLT}}^{(s)} \mathbf{P}_{\text{LDA}}^{(s)} \mathbf{x}_t^{(s)}, \quad \text{where } s = A, V, \quad (2)$$

where matrices $\mathbf{P}_{\text{LDA}}^{(s)}$ and $\mathbf{P}_{\text{MLLT}}^{(s)}$ are of dimensions $D_s \times d_s$ and $D_s \times D_s$, respectively. Values $n_A = 24$, $J_A = 9$, $D_A = 60$, and $n_V = 24$, $J_V = 15$, $D_V = 41$, are used [14].

2.2. Audio-visual fusion

Audio-visual integration methods can be broadly grouped into *feature fusion* algorithms, that use a *single* classifier trained on the concatenated vector of audio and visual features, or on any appropriate transformation of it [2], [4], [5], [6], [14], and into *decision fusion* techniques, that combine the two single-modality (audio- and visual-only) classifier outputs to recognize audio-visual speech [3], [5], [6].

For our speechreading system, we adopt two simple feature fusion strategies, considered in [14]. The first (*feature concatenation*) uses the concatenation of the synchronous audio and visual features as the joint bimodal feature vector

$$\mathbf{o}_t^{(\text{AV})} = [\mathbf{o}_t^{(A)\top}, \mathbf{o}_t^{(V)\top}]^\top \in \mathbb{R}^D, \quad (3)$$

where $D = D_A + D_V = 101$, here. This dimension is rather high compared to feature vectors used in typical LVCSR systems. To achieve dimensionality reduction of (3), we apply LDA, followed by MLLT, thus obtaining an alternative audio-visual feature vector,

$$\mathbf{o}_t^{(\text{HiLDA})} = \mathbf{P}_{\text{MLLT}}^{(\text{AV})} \mathbf{P}_{\text{LDA}}^{(\text{AV})} \mathbf{o}_t^{(\text{AV})}. \quad (4)$$

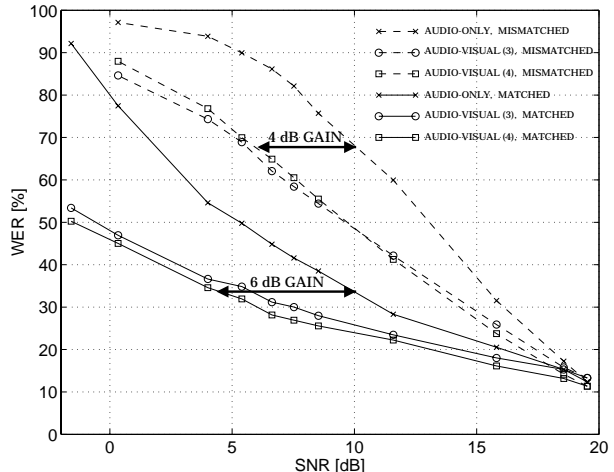


Figure 1: Test set audio-only and audio-visual word error rate (WER) using concatenative (see (3)) and HiLDA (see (4)) fusion, for both matched and unmatched training/testing, as a function of the audio channel signal-to-noise ratio (SNR).

This amounts to a second application of LDA and MLLT (see (2)), therefore this fusion method is referred to as *Hierarchical LDA*, or HiLDA. In our experiments, matrix $\mathbf{P}_{\text{LDA}}^{(\text{AV})}$ is of size 60×101 , giving rise to 60-dimensional HiLDA features. For both feature fusion algorithms (3) and (4), we model the generation process of the sequence of audio-visual features by a single-stream *hidden Markov model* (HMM), with 2,808 context dependent states and 47,160 Gaussian mixture components.

3. Automatic speechreading experiments

We use the IBM ViaVoiceTM audio-visual database for our experiments [5]. It consists of full-face frontal video and audio of 290 subjects, uttering continuous, read speech, with mostly verbalized punctuation and a 10,400 word vocabulary. The database video is sized at 704×480 pixels, interlaced, MPEG2 encoded, captured in color at a rate of 30 Hz (60 fields per second) are available at a resolution of 240 lines). High quality wideband audio is synchronously collected with the video at a rate of 16 kHz in an office environment at 19.5 dB SNR. Approximately 37.5 hours of data are used in speaker-independent audio-visual ASR experiments, partitioned into a *training* set that contains 35 hours of data (17,111 utterances) from 239 subjects, used for HMM parameter estimation, and a 2.5-hour *test* set (1,038 utterances) from 26 additional subjects, provided for HMM evaluation.

To study the benefit of the visual modality to the automatic speechreading system described in Section 2, we consider both the original database clean audio (19.5 dB SNR), as well as a number of progressively noisier audio conditions (down to -1.6 dB SNR), with the audio channel artificially degraded by additive, non-stationary, wideband, speech-babble noise. For each audio condition, we use the training set data to first estimate matrices $\mathbf{P}_{\text{LDA}}^{(s)}$, $\mathbf{P}_{\text{MLLT}}^{(s)}$, for $s = A, V$, and $\mathbf{P}_{\text{LDA}}^{(\text{AV})}$, $\mathbf{P}_{\text{MLLT}}^{(\text{AV})}$ (see (2), (4)), and subsequently train audio-only and audio-visual HMMs using concatenative (see (3)) and HiLDA (see (4)) feature fusion [14]. We then evaluate the resulting HMM performance on the test set (*matched* training/testing scenario). The results are depicted in Fig. 1 (solid lines). Notice that in all conditions considered, HiLDA features significantly outperform concatenated audio-visual features. At 19.5 dB SNR, the visual modal-

ity reduces the WER by 9% relative, when HiLDA is used (from 12.4% audio-only to a 11.3% audio-visual WER), however the WER degrades to 13.3%, when concatenative feature fusion is employed. For the 8.5 dB SNR condition considered in [14], the HiLDA obtained WER improvement is 34% relative (from 38.5% to a 25.6% WER). These results outperform the ones reported in [14], due to better trained models and the use of full audio-visual decoding, instead of audio lattice rescoring with audio-visual models. As expected, the benefit of including the visual modality increases with decreasing SNR, reaching a relative 46% at -1.6 dB (from 92.2% to a 50.2% WER), by means of HiLDA feature fusion. A useful indicator of this benefit is the effective SNR gain, measured in this paper with reference to the audio-only WER at 10 dB. To compute this gain, we need to consider the SNR value where the audio-visual WER equals the reference audio-only WER (see Fig. 1); for HiLDA fusion, the gain equals approximately 6 dB, whereas for concatenative fusion, the effective SNR gain drops to 4 dB.

In Fig. 1, we also depict audio-only and audio-visual performance for *mismatched* training/testing (dash-lines). In this scenario, the audio-only and audio-visual HMMs trained at the original database audio channel condition (19.5 dB) are evaluated at all other SNRs. As expected, the performance of all HMMs degrades, compared to the matched trained models, however, both mismatched audio-visual systems outperform mismatched audio-only ASR. Not surprisingly, as the mismatch increases, HiLDA fusion becomes worse than concatenative feature fusion. The effective SNR gain drops to about 4 dB for HiLDA and to 4.5 dB for concatenative feature fusion.

In Section 5, when comparing audio-visual ASR and human performance, we will be using HMMs trained in the matched condition. In this case, HiLDA is superior, therefore we will be only reporting HiLDA based ASR results.

4. Human speech perception experiments

Similarly to the automatic speechreading experiments reported in Section 3, we consider utterances from the IBM ViaVoiceTM audio-visual database, with their audio channel artificially corrupted by additive speech-babble noise at various levels of SNR. A small number of human listeners are then presented with the audio-only, as well as the audio-visual stimuli containing the speaker's full frontal face, and they are asked in each case to transcribe the utterances. The audio-only and audio-visual WERs for the entire pool of transcribed sentences are then computed at each SNR condition considered (see Fig. 2).

In more detail, 50 database sequences from a *single* speaker are chosen for presentation to the human listeners, and their transcriptions are verified for accuracy (recall that the database consists of read speech, hence the original transcriptions are available). Subsequently, the audio channel of each sequence is demultiplexed from the MPEG2 video file, contaminated by additive speech-babble noise at various SNRs, thus giving rise to the audio-only stimulus, and then multiplexed back into the video sequence, thus providing the audio-visual stimulus at the desired SNR level. Six audio conditions are considered in our experiments, namely the original clean database audio at 19.5 dB, and five noisy conditions corresponding to SNRs of approximately 14.5, 13.5, 9.0, 5.0, and -0.5 dB (computed over the sequences presented to the listeners at each condition).

Initially, we consider human speech perception at the -0.5 dB condition. Fourteen human listeners are presented with up to 15 sequences each, chosen at random from the available pool of 50 database utterances. Each subject listens first to up to three repetitions of the noisy audio-only sequence and is asked

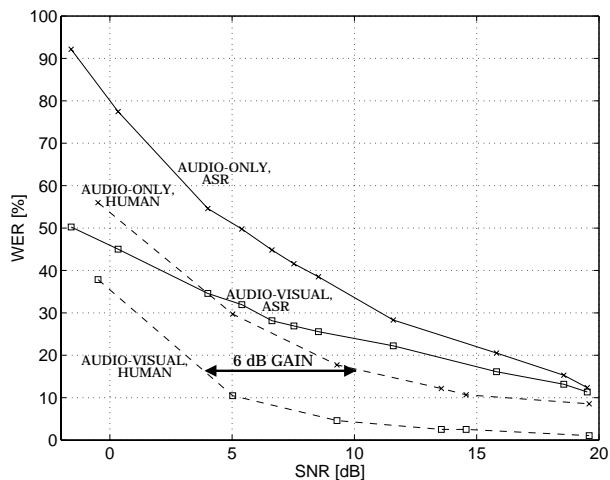


Figure 2: Automatic recognition and human perception of audio-only and audio-visual speech for various SNRs. HiLDA feature fusion and matched training are used in the former.

to transcribe it on paper. Next, the subject is asked to transcribe the utterance again, but after being presented with up to three repetitions of the noisy audio-visual sequence. A total of 194 audio-only and their 194 corresponding audio-visual sequences are transcribed at the -0.5 dB SNR condition by the 14 subjects. The transcriptions are corrected for obvious spelling errors, entered into a computer, and compared to the actual utterance transcriptions to compute the human word error rate (deletions, insertions, and substitutions are taken into account). For the 194 sequences transcribed, the human audio-only WER is 56.0% (per-subject WER varies from 45.9% to 81.7%), whereas the audio-visual WER is reduced to 37.8% (varying between 23.0% and 70.4%). All 14 subjects improve speech intelligibility using the visual modality, by as much as 57% relative, and by a cumulative 33% relative (see also Fig. 2).

Next, we consider human speech perception at the remaining five SNR conditions within the [5.0, 19.5] dB interval. For this study, we use 10 human listeners, and 30 utterances, chosen out of the original set of 50, which we partition into six 5-utterance blocks. We then assign two such 5-utterance blocks to each of our 10 subjects. Each subject first listens to the audio-only sequences of the first assigned block, for each of the five audio conditions, moving from lower to higher SNRs.¹ Next, each subject is shown the audio-visual sequences of the second assigned block, again at all five SNRs. Each stimulus is presented up to three times, and the subject is asked to transcribe the utterance at each condition. The resulting human audio-only and audio-visual WERs at each SNR are depicted in Fig. 2 (dash-lines). Clearly, the visual modality significantly helps human perception for all audio conditions considered. Notice, that an effective SNR gain of 6 dB is observed compared to the audio-only performance at 10 dB.

5. Machine versus human performance

A comparison of automatic and human perception of audio-visual speech is depicted in Fig. 2. Surprisingly, audio-only human WER in the clean condition is quite close to machine audio-only WER. This is unexpected, as human WER is known

¹The sequence of stimulus presentation is designed to expose the listener to the less informative condition first, and, subsequently, to progressively more information.

to be typically an order of magnitude less than machine WER, for a large number of audio-only recognition tasks [12]. This discrepancy is possibly due to the complex lexical content of the sequences used in the human perception experiment (many relatively unknown proper names and little linguistic context), the fact that 30% of the human listeners used are non-native speakers, and possible side-effects from transcribing lower SNR utterances first. For noisy audio, human WER degrades more slowly than in ASR (notice the use of FMN and matched training for the latter). This is a well-documented fact in the literature [12], where typically less extreme noise conditions than speech-babble are considered.

When adding the visual modality, both humans and machines benefit in recognizing speech. The effective SNR gains for both are approximately equal to 6 dB, when compared to the audio-only performance at 10 dB. However, our experiments demonstrate that the relative gains in improved WER differ at low and high SNRs. Humans make good use of the visual modality at both high and medium SNRs, as indicated by the high relative WER reduction in these conditions. However, for extreme noise conditions, such as speech-babble at 0 dB, the relative gain diminishes. In contrast, the visual modality benefit to automatic speech recognition is minimal at high SNRs but it continuously increases with decreasing SNR. Interestingly, at approximately 4 dB, the automatic speechreading system outperforms audio-only human speech perception. Fig. 2 indicates that possibly the automatic speechreading system could even achieve a lower WER than audio-visual human perception at very low (sufficiently negative) SNRs.

A word of caution is warranted in these comparisons, as our human speech perception experimental protocol has a number of shortcomings. Indeed, we considered single-speaker sequences for transcription by a small only number of human listeners. More subjects in both sides of the experiment (database speakers and human listeners) will benefit the generalization and significance of our observations [17]. A more careful selection of the database sequences to avoid “out-of-vocabulary” words for humans, such as obscure proper names, will help boost human recognition. Finally, motivating human listeners to continuously pay attention, by for example rewarding high transcription accuracy, might also be beneficial.

6. Conclusions and future work

We presented a comparison of machine and human performance when recognizing large-vocabulary, continuous audio-visual speech. We first considered audio-only and audio-visual ASR at a variety of SNRs, using a large, speaker-independent audio-visual database, suitable for LVCSR. Audio-visual results were obtained by means of an automatic speechreading system with a pixel based visual front end and audio-visual feature fusion based on hierarchical LDA. This fusion algorithm was shown to outperform plain feature concatenation for matched training/testing. Next, we considered a simple human speech perception experiment, where a small number of subjects were asked to transcribe database audio-only and audio-visual utterances, at various SNRs. The visual modality was demonstrated to significantly improve recognition for both the machine and humans, amounting to an effective SNR gain of approximately 6 dB, at the audio SNR reference of 10 dB. However the patterns of the visual benefit differ between humans and machines at low and high SNRs. Our experiments indicate that the automatic system benefits the most at low SNRs, with its relative WER improvement continuously increasing with decreasing SNR, reaching up to 46% relative at -1.6 dB. In contrast,

humans seem to benefit relatively more at high and medium SNRs. Finally, at a sufficiently low SNR level (below 4 dB), audio-visual automatic recognition outperforms audio-only human perception, thus achieving “super-human” performance.

Of course, the human speech perception study in this work is somewhat limited, for a number of reasons discussed in Section 5. In the near future, we plan to correct these shortcomings, and to also compare automatic and human recognition of audio-visual speech in additional domains, such as small-vocabulary connected-word tasks.

7. References

- [1] Stork, D.G. and Hennecke, M.E. eds., *Speechreading by Humans and Machines*, Springer, Berlin, 1996.
- [2] Teissier, P., Robert-Ribes, J., Schwartz, J.-L., and Guérin-Dugué, A., “Comparing models for audiovisual fusion in a noisy-vowel recognition task,” *IEEE Trans. Speech Audio Process.*, vol. 7, pp. 629–642, 1999.
- [3] Dupont, S. and Luettin, J., “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.
- [4] Chen, T., “Audiovisual speech processing. Lip reading and lip synchronization,” *IEEE Signal Process. Mag.*, vol. 18, pp. 9–21, 2001.
- [5] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J., “Audio-visual speech recognition,” *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000 (http://www.clsp.jhu.edu/ws2000/final_reports/avsr/).
- [6] Hennecke, M.E., Stork, D.G., and Prasad, K.V., “Visionary speech: Looking ahead to practical speechreading systems,” in [1], pp. 331–349, 1996.
- [7] Campbell, R., Dodd, B., and Burnham, D. eds., *Hearing by Eye II*, Psychology Press Ltd. Publishers, Hove, 1998.
- [8] Sumby, W.H. and Pollack, I., “Visual contribution to speech intelligibility in noise,” *J. Acoust. Soc. America*, vol. 26, pp. 212–215, 1954.
- [9] McGurk, H. and MacDonald, J.W., “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.
- [10] Summerfield, A.Q., “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lip-Reading*, Dodd, B. and Campbell, R. eds., Lawrence Erlbaum Associates, Hillsdale, pp. 97–113, 1987.
- [11] Robert-Ribes, J., Piquemal, M., Schwartz, J.-L., and Escudier, P., “Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition,” in [1], pp. 193–210, 1996.
- [12] Lippmann, R.P., “Speech recognition by machines and humans,” *Speech Communication*, vol. 22, pp. 1–15, 1997.
- [13] Bernstein, L.E. and Auer Jr., E.T., “Word Recognition in Speechreading,” in [1], pp. 17–26, 1996.
- [14] Potamianos, G., Luettin, J., and Neti, C., “Hierarchical discriminant features for audio-visual LVCSR,” *Proc. Int. Conf. Acoust. Speech Signal Process.* (In Press), 2001.
- [15] Summerfield, Q., “Use of visual information for phonetic perception,” *Phonetica*, vol. 36, pp. 314–331, 1979.
- [16] Summerfield, Q., MacLeod, A., McGrath, M., and Brooke, M., “Lips, teeth, and the benefits of lipreading,” in *Handbook of Research on Face Processing*, Young, A.W. and Ellis, H.D. eds., Elsevier Science Publishers, Amsterdam, pp. 223–233, 1989.
- [17] Kricos, P.B., “Differences in visual intelligibility across talkers,” in [1], pp. 43–53, 1996.