15

Adaptive Decision Fusion for Audio-Visual Speech Recognition

Jong-Seok Lee[1] and Cheol Hoon Park[2]
[1] *Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL)*
[2]*School of Electrical Engineering and Computer Science, KAIST*
[1]*Switzerland*
[2]*Korea*

1. Introduction

While automatic speech recognition technologies have been successfully applied to real-world applications, there still exist several problems which need to be solved for wider application of the technologies. One of such problems is noise-robustness of recognition performance; although a speech recognition system can produce high accuracy in quiet conditions, its performance tends to be significantly degraded under presence of background noise which is usually inevitable in most of the real-world applications.

Recently, audio-visual speech recognition (AVSR), in which visual speech information (i.e., lip movements) is used together with acoustic one for recognition, has received attention as a solution of this problem. Since the visual signal is not influenced by acoustic noise, it can be used as a powerful source for compensating for performance degradation of acoustic-only speech recognition in noisy conditions. Figure 1 shows the general procedure of AVSR: First, the acoustic and the visual signals are recorded by a microphone and a camera, respectively. Then, salient and compact features are extracted from each signal. Finally, the two modalities are integrated for recognition of the given speech.

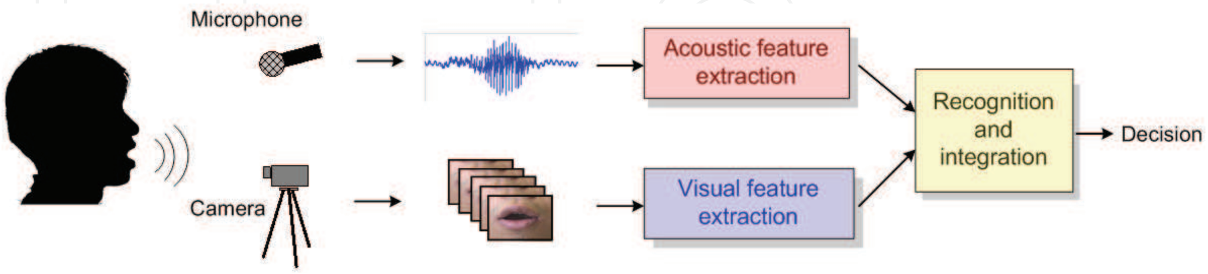

Microphone
Acoustic feature extraction
Recognition and integration
Decision
Camera
Visual feature extraction

Fig. 1. General procedure of audio-visual speech recognition

In this chapter, we focus on the problem of audio-visual information fusion in AVSR, i.e., how to combine the two modalities effectively, which is an important issue for noise-robust AVSR. A good method of audio-visual fusion should exploit complementary characteristics of the two modalities efficiently so that we can obtain robust recognition performance over various noisy environments.

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

First, we give a review of methods for information fusion in AVSR. We present biological and psychological backgrounds of audio-visual information fusion. Then, we discuss existing fusion methods. In general, we can categorize such methods into two broad classes: feature fusion (or early integration) and decision fusion (or late integration). In feature fusion, the features from the two information sources are concatenated first and then the combined features are fed into a recognizer. In decision fusion, the features of each modality are used for recognition separately and the outputs of the two recognizers are integrated for the final recognition result. Each approach has its own advantages and disadvantages, which are explained and compared in detail in this chapter.

Second, we present an adaptive fusion method based on the decision fusion approach. Between the two fusion approaches explained above, it has been shown that decision fusion is more preferable for implementing noise-robust AVSR systems than feature fusion. In order to construct a noise-robust AVSR system adopting decision fusion, it is necessary to measure relative reliabilities of the two modalities for given speech data and to control the amounts of the contribution of the modalities according to the measured reliabilities. Such an adaptive weighting scheme enables us to obtain robust recognition performance consistently over diverse noise conditions. We compare various definitions of the reliability measure which have been suggested in previous researches. Then, we introduce a neural network-based method which is effective for generating appropriate weights for given audio-visual speech data of unknown noise conditions and thereby producing robust recognition results in a wide range of operating conditions.

2. Audio-visual speech recognition and information fusion

The ultimate goal of the AVSR technology would be to construct a recognizer whose performance is comparable to that of humans. Thus, understanding how humans perceive audio-visual speech will be helpful for constructing AVSR systems showing good performance. In this section, we review theories of modality integration in humans' bimodal speech perception in the viewpoint of biology and psychology, and presents approaches of information fusion for AVSR.

2.1 Bimodal nature of speech perception

The process of humans' speech production is intrinsically bimodal: The configuration of the tongue, the jaw, the teeth and the lips determines which specific sound is produced. Many of such articulatory movements are visible. Therefore, the mechanism of humans' speech perception is also bimodal. In a face-to-face conversation, we listen to what others say and, at the same time, observe their lip movements, facial expressions, and gestures. Especially, if we have a problem in listening due to environmental noise, the visual information plays an important role for speech understanding (Ross et al., 2007). Even in the clean condition speech recognition performance is improved when the talking face is visible (Arnold & Hill, 2001). Also, it is well-known that hearing-impaired people often have good lipreading skills.

There exist many researches proving the bimodality of speech perception and showing interesting results of audio-visual interaction due to the bimodality: The McGurk effect demonstrated the bimodality of the humans' speech perception by showing that, when the acoustic and the visual speech is incongruent, listeners recognize the given speech as a sound which is neither the acoustic nor the visual speech (McGurk & MacDonald, 1976). It was shown that many phonemes which are acoustically confusable are easily distinguished using visual information (for example, /b/ and /g/) (Summerfield, 1987). Psychological experiments showed that seeing speakers' lips enhances the ability to detect speech in noise by decreasing auditory detection threshold of speech in comparison to the audio-only case, which is called "bimodal coherence masking protection" meaning that the visual signal acts as a cosignal assisting auditory target detection (Grant & Seitz, 2000; Kim & Davis, 2004). Such improvement is based on the correlations between the acoustic signal and the visible articulatory movement. Moreover, the enhanced sensitivity improves the ability to understand speech (Grant & Seitz, 2000; Schwartz et al., 2004).

A neurological analysis of the human brain shows an evidence of humans' multimodal information processing capability (Sharma et al., 1998): When different senses reach the brain, the sensory signals converge to the same area in the superior colliculus. A large portion of neurons leaving the superior colliculus are multisensory. In this context, a neurological model of sensor fusion has been proposed, in which sensory neurons coming from individual sensors are fused in the superior colliculus (Stein & Meredith, 1993). Also, it has been shown through positron emission tomography (PET) experiments that audio-visual speech perception yields increased activity in multisensory association areas such as superior temporal sulcus and inferior parietal lobule (Macaluso et al., 2004). Even silent lipreading activates the primary auditory cortex, which is shown by neuroimaging researches (Calvert et al., 1997; Pekkola et al., 2005; Ruytjens et al., 2007).

The nature of humans' perception demonstrates a statistical advantage of bimodality: When humans have estimates of an environmental property from two different sensory systems, any of which is possibly corrupted by noise, they combine the two signals in the statistically optimal way so that the variance of the estimates for the property is minimized after integration. More specifically, the integrated estimate is given by the maximum likelihood rule in which the two unimodal estimates are integrated by a weighted sum with each weight inversely proportional to the variance of the estimate by the corresponding modality (Ernest & Banks, 2002).

The advantage of utilizing the acoustic and the visual modalities for human speech understanding comes from the following two factors. First, there exists "complementarity" of the two modalities: The two pronunciations /b/ and /p/ are easily distinguishable with the acoustic signal, but not with the visual signal; on the other hand, the pronunciations /b/ and /g/ can be easily distinguished visually, but not acoustically (Summerfield, 1987). From the analysis of French vowel identification experiments, it has been shown that speech features such as height (e.g., /i/ vs. /o/) and front-back (e.g., /y/ vs. /u/) are transmitted robustly by the acoustic channel, whereas some other features such as rounding (e.g., /i/ vs. /y/) are transmitted well by the visual channel (Robert-Ribes et al., 1998). Second, the two modalities produce "synergy.": Performance of audio-visual speech perception can outperform those of acoustic-only and visual-only perception for diverse noise conditions (Benoît et al., 1994).

2.2 Theories of bimodal speech perception

While the bimodality of speech perception has been widely demonstrated as shown above, its mechanism has not been clearly understood yet because it would require wide and deep psychological and biological understanding about the mechanisms of sensory signal processing, high-level information processing, language perception, memory, etc. In this subsection, we introduce some existing psychological theories to explain how humans perform bimodal speech perception, some of which conflict with each other.

There exists a claim that visual speech is secondary to acoustic speech and affects perception only when the acoustic speech is not intelligible (Sekiyama & Tohkura, 1993). However, the McGurk effect is a counterexample of this claim; the effect is observed even when the acoustic speech is not corrupted by noise and clearly intelligible.

The direct identification model by Summerfield is an extension of Klatt's lexical-access-from-spectra model (Klatt, 1979) to a lexical-access-from-spectra-and-face-parameters model (Summerfield, 1987). The model assumes that the bimodal inputs are processed by a single classifier. A psychophysical model based on the direct identification has been derived from the signal detection theory for predicting the confusions of audio-visual consonants when the acoustic and the visual stimuli are presented separately (Braida, 1991).

The motor theory assumes that listeners recover the neuromotor commands to the articulators (referred to as "intended gestures") from the acoustic input (Liberman & Mattingly, 1985). The space of the intended gestures, which is neither acoustic nor visual, becomes a common space where the two signals are projected and integrated. A motivation of this theory is the belief that the objects of speech perception must be invariant with respect to phonemes or features, which can be achieved only by neuromotor commands. It was argued that the motor theory has a difficulty in explaining the influence of higher-order linguistic context (Massaro, 1999).

The direct realist theory also claims that the objects of speech perception are articulatory rather than acoustic events. However, in this theory the articulatory objects are actual, phonetically structured vocal tract movements or gestures rather than the neuromotor commands (Fowler, 1986).

The TRACE model is an interactive activation model in which excitatory and inhibitory interactions among simple processing units are involved in information processing (McClelland & Elman, 1986). There are three levels of units, namely, feature, phoneme and word, which compose of a bidirectional information processing channel: First, features activate phonemes, and phonemes activate words. And, activation of some units at a level inhibits other units of the same level. Second, activation of higher level units activates their lower level units; for example, a word containing the /a/ phoneme activates that phoneme. Visual features can be added to the TRACE of the acoustic modality, which produces a model in which separate feature evaluation of acoustic and visual information sources is performed (Campbell, 1988).

The fuzzy logical model of perception (FLMP) is one of the most appealing theories for humans' bimodal speech perception. It assumes perceiving speech is fundamentally a pattern recognition problem, where information processing is conducted with probabilities as in Bayesian analysis. In this model, the perception process consists of the three stages which are successive but overlapping, as illustrated in Figure 2 (Massaro, 1987; Massaro, 1998; Massaro, 1999): First, in the evaluation stage, each source of information is evaluated to produce continuous psychological values for all categorical alternatives (i.e., speech classes). Here, independent evaluation of each information source is a central assumption of the FLMP. The psychological values indicate the degrees of match between the sensory information and the prototype descriptions of features in memory, which are analogous to the fuzzy truth values in the fuzzy set theory. Second, the integration stage combines these to produce an overall degree of support for each alternative, which includes multiplication of the supports of the modalities. Third, the decision stage maps the outputs of integration into some response alternative which can be either a discrete decision or a likelihood of a given response.

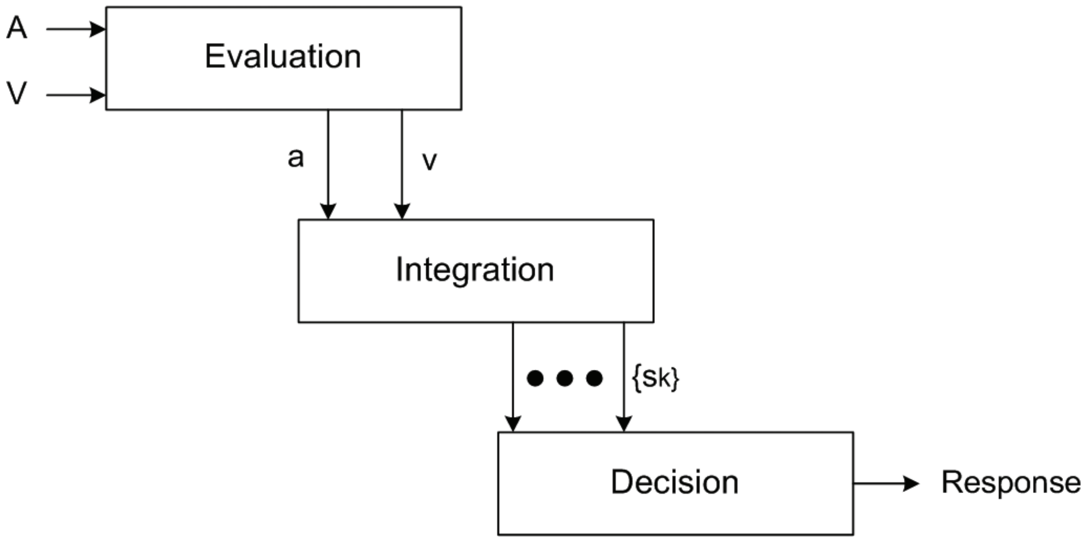


Fig. 2. Illustration of the processes of perception in FLMP. *A* and *V* represent acoustic and visual information, respectively. *a* and *v* are psychological values produced by evaluation of *A* and *V*, respectively. s_k is the overall degree of support for the speech alternative *k*.

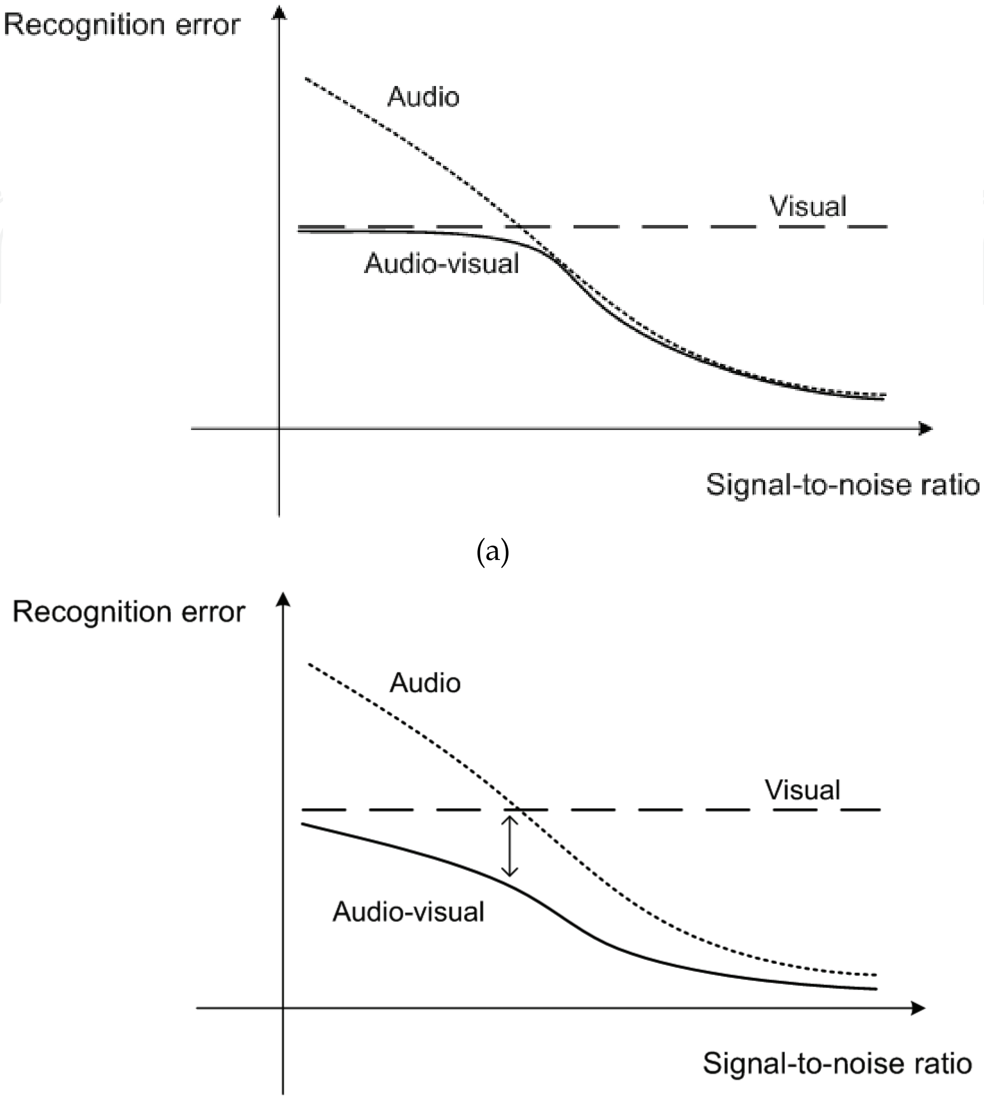
It is worth mentioning about the validity of the assumption that there is no interaction between the modalities. Some researchers have argued that interaction between the acoustic and the visual modalities occurs, but it has also argued that very little interaction occurs in human brains (Massaro & Stork, 1998). In addition, the model seems to successfully explain several perceptual phenomena and be broadening its domain, for example, individual differences in speech perception, cross-linguistic differences, distinction between information and information-processing. Also, it has been shown that the FLMP gives better description of various psychological experiment results than other integration models (Massaro, 1999).

2.3 Approaches for information fusion in AVSR

The primary challenge in AVSR is to obtain the performance which is equal to or better than the performance of any modality for various noise conditions. When the noise level is low, the acoustic modality performs better than the visual one and, thus, the audio-visual recognition performance should be at least as good as that of the acoustic speech recognition. When the noise level is high and the visual recognition performance is better than the acoustic one, the integrated recognition performance should be at least the same to or better than the performance of the visual-only recognition.

Besides, we expect the synergy effect of the two modalities by using AVSR systems. Thus, the goal of the second challenge in the use of audio-visual information for speech recognition is to improve the recognition performance with as a high synergy of the modalities as possible.

These two challenges are illustrated in Figure 3. The audio-visual information fusion process is an important issue causing the gap of the audio-visual recognition performance of the two cases in the figure. Combining the two modalities should take full advantage of the modalities so that the integrated system shows a high synergy effect for a wide range of noise conditions. On the contrary, when the fusion is not performed appropriately, we cannot expect complementarity and synergy of the two information sources and, moreover, the integrated recognition performance may be even inferior to that of any of the unimodal systems, which is called “attenuating fusion” or “catastrophic fusion” (Chibelushi et al., 2002).

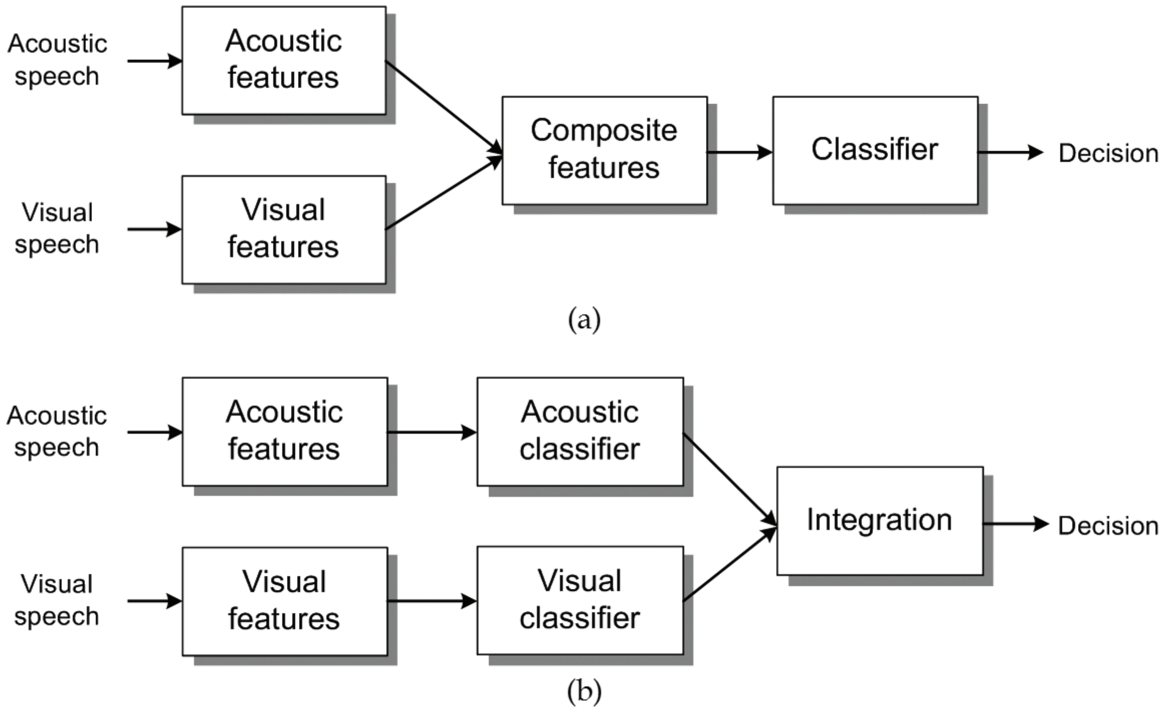


(a)

(b)

Fig. 3. Two challenges of AVSR. (a) The integrated performance is at least that of the modality showing better performance for each noise level. (b) The integrated recognition system shows the synergy effect.

In general, we can categorize methods of audio-visual information fusion into two broad categories: feature fusion (or early integration) and decision fusion (or late integration), which are shown in Figure 4. In the former approach, the features of the two modalities are concatenated to form a composite feature vector, which is inputted to the classifier for recognition. In the latter approach, the features of each modality are used for recognition separately and, then, the outputs of the two classifiers are combined for the final recognition result. Note that the decision fusion approach shares a similarity with the FLMP explained in the previous subsection in that both are based on the assumption of class-conditional independence, i.e., the two information sources are evaluated (or recognized) independently.



(a)

(b)

Fig. 4. Models for integrating acoustic and visual information. (a) Feature fusion. (b) Decision fusion.

Although which approach is more preferable is still arguable, there are some advantages of the decision fusion approach in implementing a noise-robust AVSR system. First, in the decision fusion approach it is relatively easy to employ an adaptive weighting scheme for controlling the amounts of the contributions of the two modalities to the final recognition according to the noise level of the speech, which is because the acoustic and the visual signals are processed independently. Such an adaptive scheme facilitates achieving the main goal of AVSR, i.e., noise-robustness of recognition over various noise conditions, by utilizing the complementary nature of the modalities effectively. Second, the decision fusion allows flexible modelling of the temporal coherence of the two information streams, whereas the feature fusion assumes a perfect synchrony between the acoustic and the visual feature sequences. It is known that there exists an asynchronous characteristic between the acoustic and the visual speech: The lips and the tongue sometimes start to move up to several hundred milliseconds before the acoustic speech signal (Benoît, 2000). In addition, there exists an "intersensory synchrony window" during which the human audio-visual speech perception performance is not degraded for desynchronized audio-visual speech (Conrey & Pisoni, 2006). Third, while it is required to train a whole new recognizer for constructing a feature fusion-based AVSR system, a decision fusion-based one can be organized by using existing unimodal systems. Fourth, in the feature fusion approach the combination of the acoustic and the visual features, which is a higher dimensional feature vector, is processed by a recognizer and, thus, the number of free parameters of the recognizer becomes large. Therefore, we need more training data to train the recognizer sufficiently in the feature fusion approach. To alleviate this, dimensionality reduction methods such as principal component analysis or linear discriminant analysis can be additionally used after the feature concatenation.

3. Decision fusion with adaptive weighting scheme

The dominant paradigm for acoustic and visual speech recognition is the hidden Markov model (HMM) (Rabiner, 1989). We train an HMM to construct a model for the acoustic or visual utterance of a speech class. And, the set of HMMs for all speech classes form a speech classifier.

As discussed in the previous section, the decision fusion approach is a good choice for designing a noise-robust AVSR system. Decision fusion in HMM-based AVSR systems is performed by utilizing the outputs of the acoustic and the visual HMMs for a given audio-visual speech datum. The important issue is how to implement adaptive decision fusion to obtain noise-robustness over various noise environments. To solve this, it is necessary to define the relative reliability measure of a modality (which is affected by the noise level) and determine an appropriate weight based on the measured reliabilities.

In this section, we present the principle of adaptive weighting, various definitions of the reliability measure, and a neural network-based method for obtaining proper integration weights according to the reliabilities.

3.1 Adaptive weighting

Adaptive weighting in decision fusion is performed in the following way: When the acoustic and the visual features (O_A and O_V) of a given audio-visual speech datum of unknown class are obtained, the recognized utterance class C^* is given by (Rogozan & Deléglise, 1998)

$$C^* = \arg\max_i \left\{ \gamma \log P(O_A \mid \lambda_A^i) + (1-\gamma) \log P(O_V \mid \lambda_V^i) \right\}, \tag{1}$$

where λ_A^i and λ_V^i are the acoustic and the visual HMMs for the *i*-th class, respectively, and $\log P(O_A \mid \lambda_A^i)$ and $\log P(O_V \mid \lambda_V^i)$ are their outputs (log-likelihoods). The integration weight γ determines how much the final decision relatively depends on each modality. It has a value between 0 and 1, and varies according to the amounts of noise contained in the acoustic speech. When the acoustic speech is clean, the weight should be large because recognition with the clean acoustic speech usually outperforms that with the visual speech; on the other hand, when the acoustic speech contains much noise, the weight should be sufficiently small. Therefore, for noise-robust recognition performance over various noise conditions, it is important to automatically determine an appropriate value of the weight according to the noise condition of the given speech signal.

3.2 Reliability measures

The reliability of each modality can be measured from the outputs of the corresponding HMMs. When the acoustic speech does not contain any noise, there are large differences between the acoustic HMMs' outputs. The differences become small when the acoustic speech contains noise, which reflects increased ambiguity in recognition due to the noise. This phenomenon is illustrated in Figure 5 which shows the outputs (log-likelihoods) of the HMMs for all utterance classes when a speech datum of clean or noisy condition is presented. (An utterance of the fourth class in the DIGIT database described in Section 4.1 is used for obtaining the result in the figure. For the acoustic features and the recognizer, refer to Sections 4.2 and 4.4, respectively.)

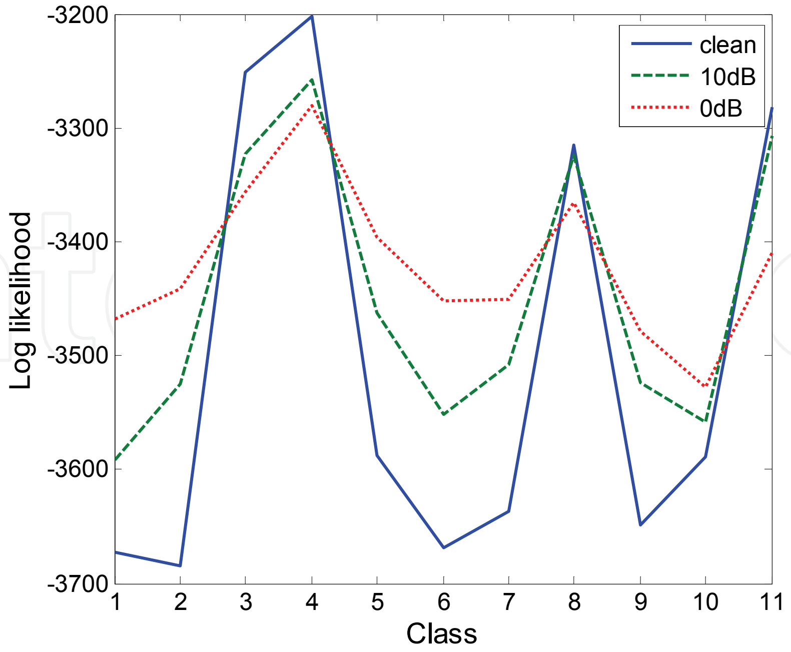


Fig. 5. Outputs of HMMs for different noise levels.

Considering this observation, we can define the reliability of a modality in various ways:

- Average absolute difference of log-likelihoods (**AbsDiff**) (Adjoudani & Benoît, 1996):

$$S = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} | L^i - L^j | , \tag{2}$$

where $L^i = \log P(O | \lambda^i)$ is the output of the HMM for the i-th class and N the number of classes being considered.

- Variance of log-likelihoods (**Var**) (Lewis & Powers, 2004):

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (L^i - \bar{L})^2 , \tag{3}$$

where $\bar{L} = \frac{1}{N} \sum_{i=1}^{N} L^i$ is the average of the outputs of the N HMMs.

- Average difference of log-likelihoods from the maximum (**DiffMax**) (Potamianos & Neti, 2000):

$$S = \frac{1}{N-1} \sum_{i=1}^{N} \left| \max_j L^j - L^i \right| , \tag{4}$$

which means the average difference between the maximum log-likelihood and the other ones.

- Inverse entropy of posterior probabilities (**InvEnt**) (Matthews et al., 1996):

$$S = \left[-\frac{1}{N} \sum_{i=1}^{N} P(C_i | O) \log P(C_i | O) \right]^{-1} , \tag{5}$$

where $P(C_i \mid O)$ is the posterior probability which is calculated by

$$P(C_i \mid O) = \frac{P(O \mid \lambda^i)}{\sum_{j=1}^{N} P(O \mid \lambda^j)}. \tag{6}$$

As the signal-to-noise ratio (SNR) value decreases, the differences of the posterior probabilities become small and the entropy increases. Thus, the inverse of the entropy is used as a measure of the reliability.

Performance of the above measures in AVSR will be compared in Section 4.

3.3 Neural network-based fusion

A neural network models the input-output mapping between the two reliabilities and the integrating weight so that it estimates the optimal integrating weights as shown in Figure 6 (Lee & Park, 2008), i.e.,

$$\hat{\gamma} = f(S_A, S_V), \tag{7}$$

where f is the function modelled by the neural network and $\hat{\gamma}$ the estimated integrating weight for the given acoustic and visual reliabilities (S_A and S_V, respectively). The universal approximation theorem of neural networks states that a feedforward neural network can model any arbitrary function with a desired error bound if the number of its hidden neurons is not limited (Hornik et al., 1989).

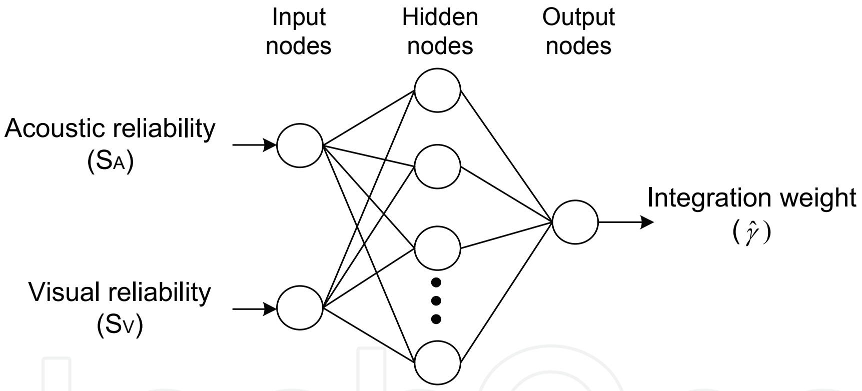

Input nodes
Hidden nodes
Output nodes
Acoustic reliability (SA)
Visual reliability (SV)
Integration weight ($\hat{\gamma}$)

Fig. 6. Neural network for estimating integration weights.

The neural network should be trained before it is used as an estimator of the integrating weight. To ensure that we obtain appropriate weights for various noise conditions by using the neural network, both clean and noisy speech data are used for training. Since it is practically impossible to use the data of all possible noise conditions, we use only speech data for a few sampled conditions. Specifically, the clean, 20 dB, 10 dB and 0 dB noisy speech data corrupted by white noise are used for training. Then, the neural network produces appropriate weights for the noise conditions which are not considered during training by its generalization capability.

Training is conducted as follows: First, we calculate the reliability of each modality for each training datum by using one of the reliability measures described in Section 3.2. Then, we obtain the integrating weights for correct recognition of the datum exhaustively; while increasing the weight from 0 to 1 by 0.01, we test whether the recognition result using the weight value is correct. Finally, the neural network is trained by using the reliabilities of the two modalities and the found weights as the training input and target pairs.

The integrating weight for correct recognition appears as an interval instead of a specific value. Figure 7 shows an example of this. It is observed that for a large SNR a large interval of the weight produces correct recognition and, as the SNR becomes small, the interval becomes small.

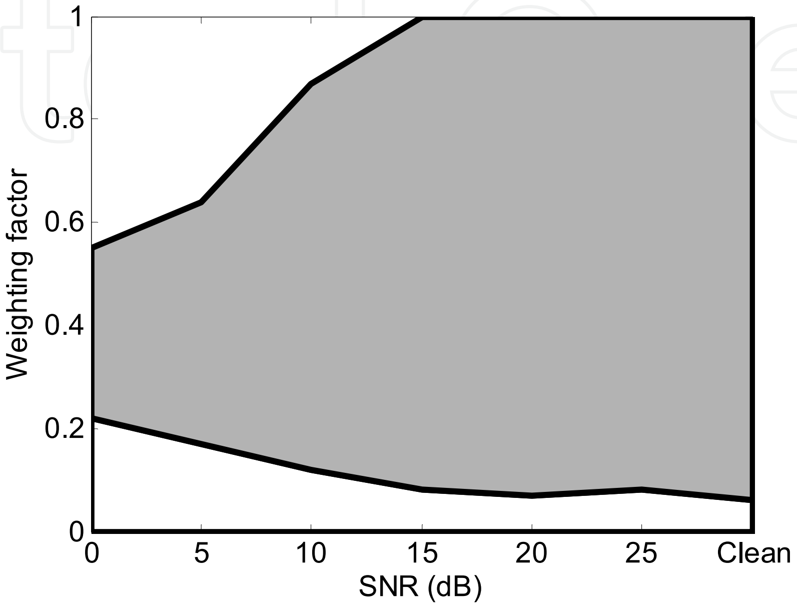


Fig. 7. Intervals of the integration weight producing correct recognition.

Therefore, the desired target for a training input vector of the neural network is given by an interval. To deal with this in training, the original error function used in the training algorithm of the neural network,

$$e(y) = t - y\,, \tag{8}$$

where t and y are the target and the output of the network, respectively, is modified as

$$e(y) = \begin{cases} \gamma_l - y & \text{for } y < \gamma_l \\ 0 & \text{for } \gamma_l \le y \le \gamma_u\,, \\ \gamma_u - y & \text{for } \gamma_u < y \end{cases} \tag{9}$$

where γ_l and γ_u are the lower and the upper bounds of the interval of the target weight value, respectively, which correspond to the boundaries of the shaded region in Figure 7.

4. Experiments

4.1 Databases

We use the two isolated word databases for experiments: the DIGIT database and the CITY database (Lee & Park, 2006). The DIGIT database contains eleven digits in Korean (including two versions of zero) and the CITY database sixteen famous Korean city names. Fifty six speakers pronounced each word three times for both databases. While a speaker was pronouncing a word, a video camera and a microphone simultaneously recorded the face region around the speaker's mouth and the acoustic speech signal, respectively. The acoustic speech was recorded at the rate of 32 kHz and downsampled to 16 kHz for feature extraction. The speaker's lip movements were recorded as a moving picture of size 720x480 pixels at the rate of 30 Hz.

The recognition experiments were conducted in a speaker-independent manner. To increase reliability of the experiments, we use the jackknife method; the data of 56 speakers are divided into four groups and we repeat the experiment with the data of the three groups (42 speakers) for training and those of the remaining group (14 speakers) for test.

For simulating various noisy conditions, we use four noise sources of the NOISEX-92 database (Varga & Steeneken, 1993): the white noise (WHT), the F-16 cockpit noise (F16), the factory noise (FAC), and the operation room noise (OPS). We add each noise to the clean acoustic speech to obtain noisy speech of various SNRs.

4.2 Acoustic feature extraction

The popular Mel-frequency cepstral coefficients (MFCCs) are extracted from the acoustic speech signal (Davis & Mermelstein, 1980). The frequency analysis of the signal is performed for each frame segmented by the Hamming window having the length of 25 ms and moving by 10 ms at a time. For each frame we perform the Fourier analysis, computation of the logarithm of the Mel-scale filterbank energy, and the discrete cosine transformation. The cepstral mean subtraction (CMS) method is applied to remove channel distortions existing in the speech data (Huang et al., 2001). As a result, we obtain 12-dimensional MFCCs, the normalized frame energy, and their temporal derivatives (i.e., delta terms).

4.3 Visual feature extraction

The visual features must contain crucial information which can discriminate between the utterance classes and, at the same time, is common across speakers having different colors of skins and lips and invariant to environmental changes such as illuminations.

In general, there are two broad categories of visual speech feature extraction: the contour-based method and the pixel-based method. The contour-based approach concentrates on identifying the lip contours. After the lip contours are tracked in the image sequences, certain measures such as the height or width of the mouth opening are used as features (Kaynak et al., 2004), or a model of the contours is built and a set of parameters describing the model configuration is used as a feature vector (Dupont & Luettin, 2000; Gurbuz et al., 2001). In the pixel-based approach, the image containing the mouth is either used directly or after some image transformations (Bregler & Konig, 1994; Lucey, 2003). Image transformation methods such as principal component analysis (PCA), discrete cosine transform and discrete wavelet transform are frequently used.

We carefully design the method of extracting the lip area and define an effective representation of the visual features derived from the extracted images of the mouth region. Our method is based on the pixel-based approach because it has advantages over the contour-based one: It does not need a complicated algorithm for accurate tracking of the lip contours and does not lose important information describing the characteristics of the oral cavity and the protrusion of lips (Matthews et al, 2001). Figure 8 summarizes the overall procedure of extracting visual features.

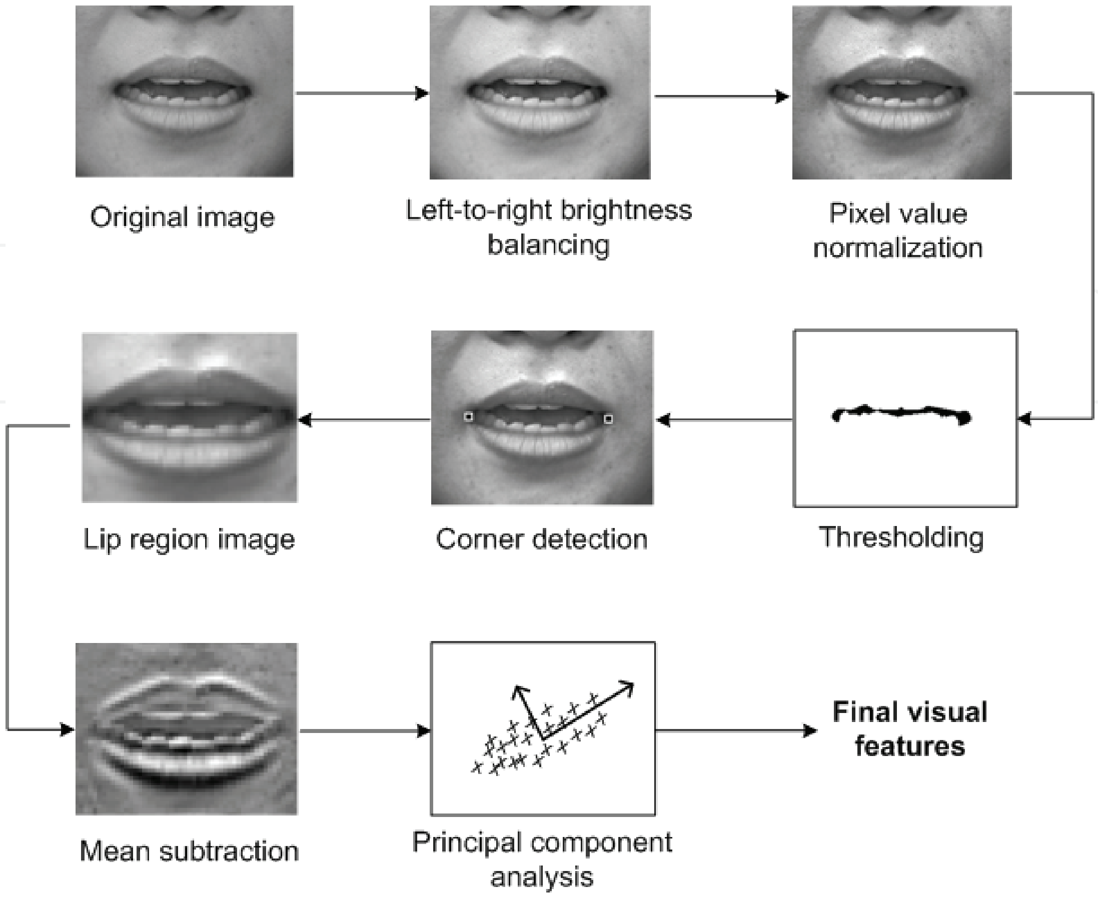


Fig. 8. Procedure of visual feature extraction.

1. We remove the brightness variation across the left and the right parts of an image so that the two mouth corners are accurately detected. We model the gradual horizontal brightness variation as the linear interpolation of the average pixel values of the left and the right small regions in the image. Then, this brightness variation is subtracted from the image in the logarithmic domain.
2. Normalization of the pixel values in the image is performed so that the pixel values of all incoming images have the same distribution characteristic. This reduces the variations of illumination conditions across recording sessions and the skin color difference across speakers. We found that the distribution of whole pixel values of the images in the database can be approximated by a Gaussian distribution. Thus, we set this Gaussian distribution as a target distribution and each image is transformed into a new image which follows the target distribution by the histogram specification technique (Gonzalez & Woods, 2001).
3. To find the mouth corners, we apply the bi-level thresholding method on the images. The thresholding is applicable for detecting mouth corners because there are always dark regions between upper and lower lips; when the mouth is open, the oral cavity appears dark, and when the mouth is closed, the boundary line between the lips appears dark. After thresholding, the left and the right end points of the dark region are the mouth corners. The mouth region is cropped based on the found corner points, so that we obtain scale- and rotation-invariant lip region images of 44x50 pixels.

4. For each pixel point, the mean value over an utterance is subtracted. Let *I(m,n,t)* be the *(m,n)*-th pixel value of the lip region image at the *t*-th frame. Then, the pixel value after mean subtraction is given by

$$J(m,n,t) = I(m,n,t) - \frac{1}{T}\sum_{t=1}^{T} I(m,n,t)\,, \tag{10}$$

   where T is the total length of the utterance. This is similar to the CMS technique in acoustic feature extraction and removes unwanted variations across image sequences due to the speakers' appearances and the different illumination conditions.

5. Finally, we apply PCA to find the main linear modes of variations and reduce the feature dimension. If we let **x** be the n_0-dimensional column vector for the pixel values of the mean-subtracted image, the n-dimensional visual feature vector **s** is given by

$$\mathbf{s} = P^T(\mathbf{x} - \bar{\mathbf{x}})\,, \tag{11}$$

   where $\bar{\mathbf{x}}$ is the mean of **x** for all training data, P is the n_0-by-n matrix whose columns are the eigenvectors for the n largest eigenvalues of the covariance matrix for all **x**'s. Here, n is much smaller than n_0(=44x50=2200) so that we obtain a compact visual feature vector. We set n to 12 in our experiment so that we obtain 12 static features for each frame. We also use the temporal derivatives of the static features as in the acoustic feature extraction.

Figure 9 shows the mean image of the extracted lip region images and the four most significant principal modes of intensity variations by ±2 standard deviations (std.) for the training data of the DIGIT database. We can see that each mode explains distinct variations

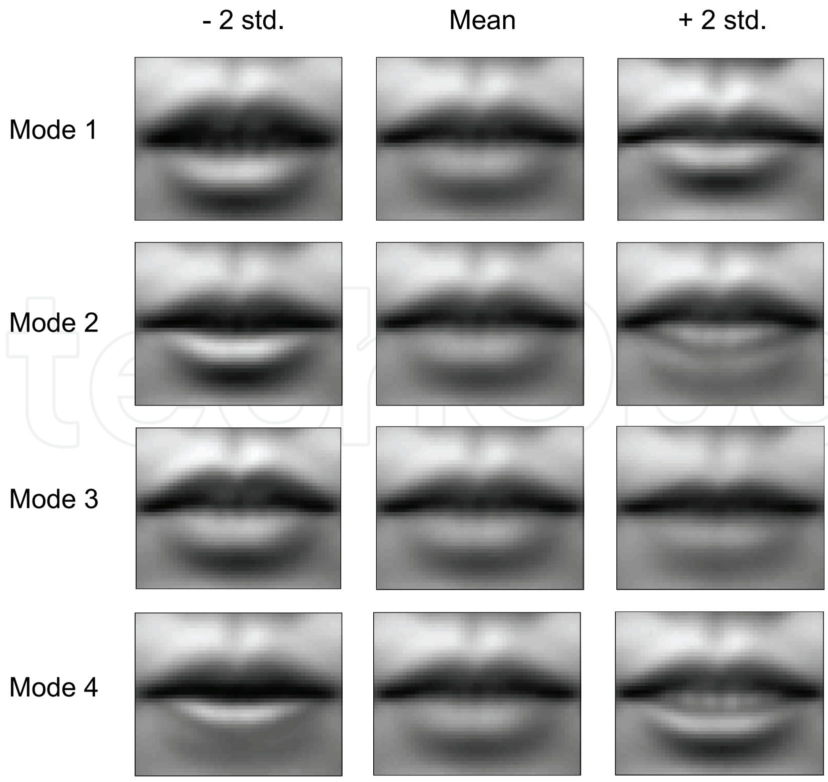


Fig. 9. First four principal modes of variations in the lip region images.

occurring in the mouth images. The first mode mainly accounts for the mouth opening. The second mode shows the protrusion of the lower lip and the visibility of the teeth. In the third mode, the protrusion of the upper lip and the changes of the shadow under the lower lip are shown. The fourth mode largely describes the visibility of the teeth.

4.4 Recognizer

The recognizer is composed of typical left-to-right continuous HMMs having Gaussian mixture models (GMMs) in each state. We use the whole-word model which is a standard approach for small vocabulary speech recognition tasks. The number of states in each HMM is set to be proportional to the number of the phonetic units of the corresponding word. The number of Gaussian functions in each GMM is set to three, which is determined experimentally. The HMMs are initialized by uniform segmentation of the training data onto the HMMs' states and iterative application of the segmental k-means algorithm. For training the HMMs, the popular Baum-Welch algorithm is used (Rabiner, 1989).

4.5 Results

First, we compare the reliability measures presented in Section 3.2. The audio-visual fusion is performed using the neural networks having five sigmoidal hidden neurons because use of more neurons did not show performance improvement. The Levenberg-Marquardt algorithm (Hagan & Menhaj, 1994), which is one of the fastest training algorithms of neural networks, is used to train the networks.

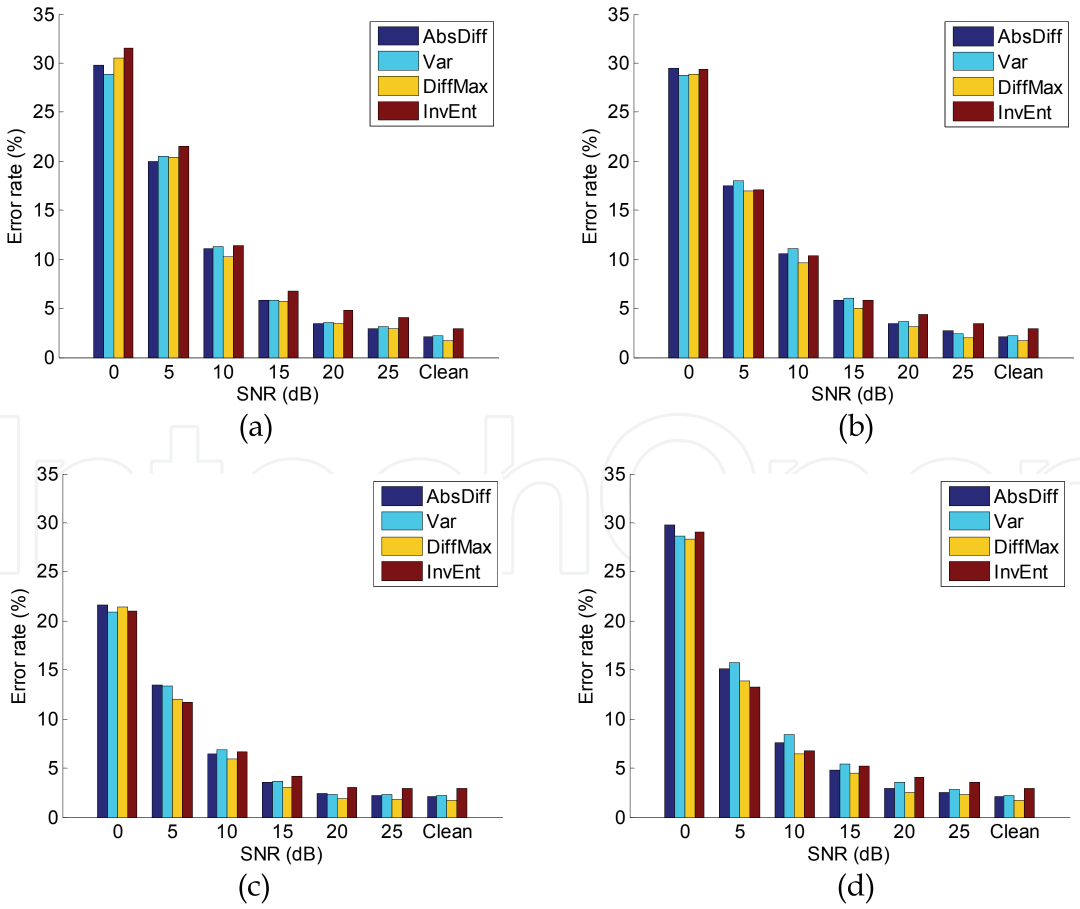
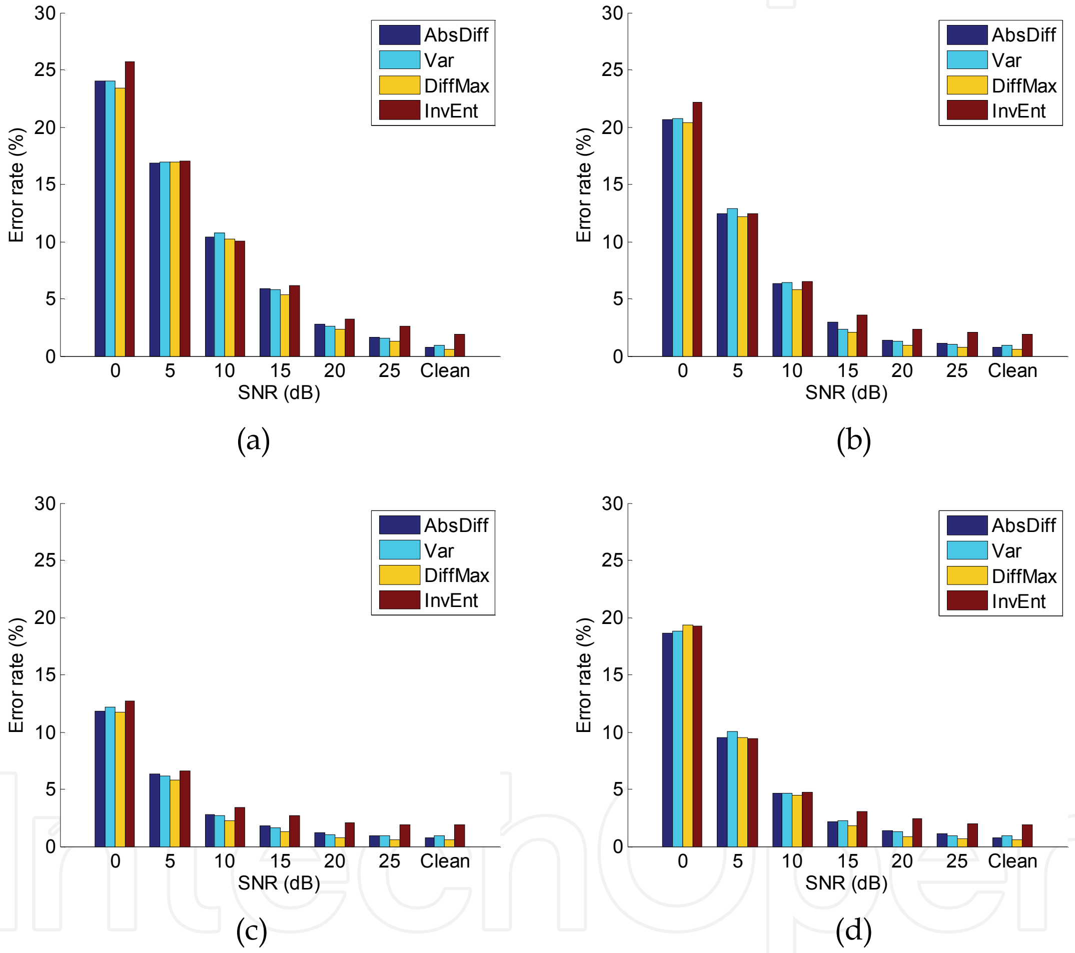


Fig. 10. Comparison of the reliability measures for the DIGIT database. (a) WHT. (b) F16. (c) FAC. (d) OPR.

Figures 10 and 11 compare the reliability measures for each database, respectively. It is observed that **DiffMax** shows the best recognition performance in an overall sense. The inferiority of **AbsDiff**, **Var** and **InvEnt** to **DiffMax** is due to their intrinsic errors in measuring reliabilities from the HMM's outputs (Lewis & Powers, 2004): Suppose that we have four classes for recognition and the HMMs' outputs are given as probabilities (e.g., [0.2, 0.4, 0.1, 0.5]). We want to get the maximum reliability when the set of the HMMs' outputs is [1, 0, 0, 0] after sorting. However, **AbsDiff** and **Var** have the maximum values when the set of the HMMs' outputs is [1, 1, 0, 0]. Also, they have the same values for [1, 0, 0, 0] and [1, 1, 1, 0], which are actually completely different cases. As for **InvEnt**, when we compare the cases of [0.1, 0.1, 0.4, 0.4] and [0.1, 0.2, 0.2, 0.5], the former has a higher value than the latter, which is the opposite of what we want.



(a) (b)

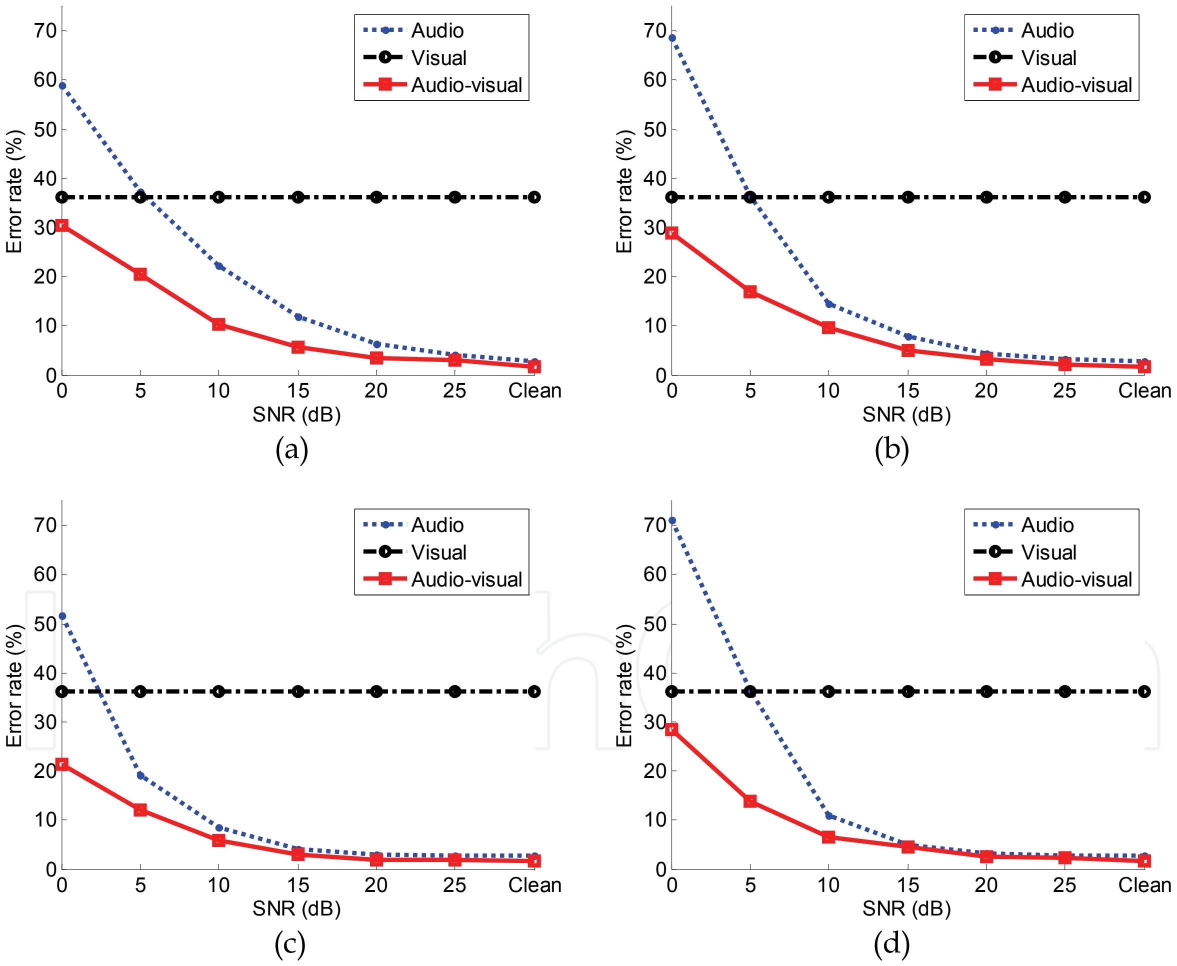
(c) (d)

Fig. 11. Comparison of the reliability measures for the CITY database. (a) WHT. (b) F16. (c) FAC. (d) OPR.

Next, we examine the unimdal and the bimodal recognition performance. Figures 12 and 13 compare the acoustic-only, the visual-only and the integrated recognition performance in error rates for the two databases, respectively. From the results, we can observe the followings:

1. The acoustic-only recognition shows nearly 100% for clean speech but, as the speech contains more noise, its performance is significantly degraded; for some noise the error rate is even higher than 70% at 0dB.

2. The error rate of the visual-only recognition is 36.1% and 22.0% for each database, respectively, which appears constant regardless of noise conditions. These values are larger than the acoustic-only recognition performance for clean speech but smaller than that for noisy speech.
3. The performance of the integrated system is at least similar to or better than that of the unimodal system. Especially, the synergy effect is prominent for 5dB~15dB. Compared to the acoustic-only recognition, relative reduction of error rates by the bimodal recognition is 39.4% and 60.4% on average for each database, respectively. For the high-noise conditions (i.e., 0dB~10dB), relative reduction of error rates is 48.4% and 66.9% for each database, respectively, which demonstrates that the noise-robustness of recognition is achieved.
4. The neural network successfully works for untrained noise conditions. For training the neural network, we used only clean speech and 20dB, 10dB and 0dB noisy speech corrupted by white noise. However, the integration is successful for the other noise levels of the same noise source and the noise conditions of the other three noise sources.



(a)

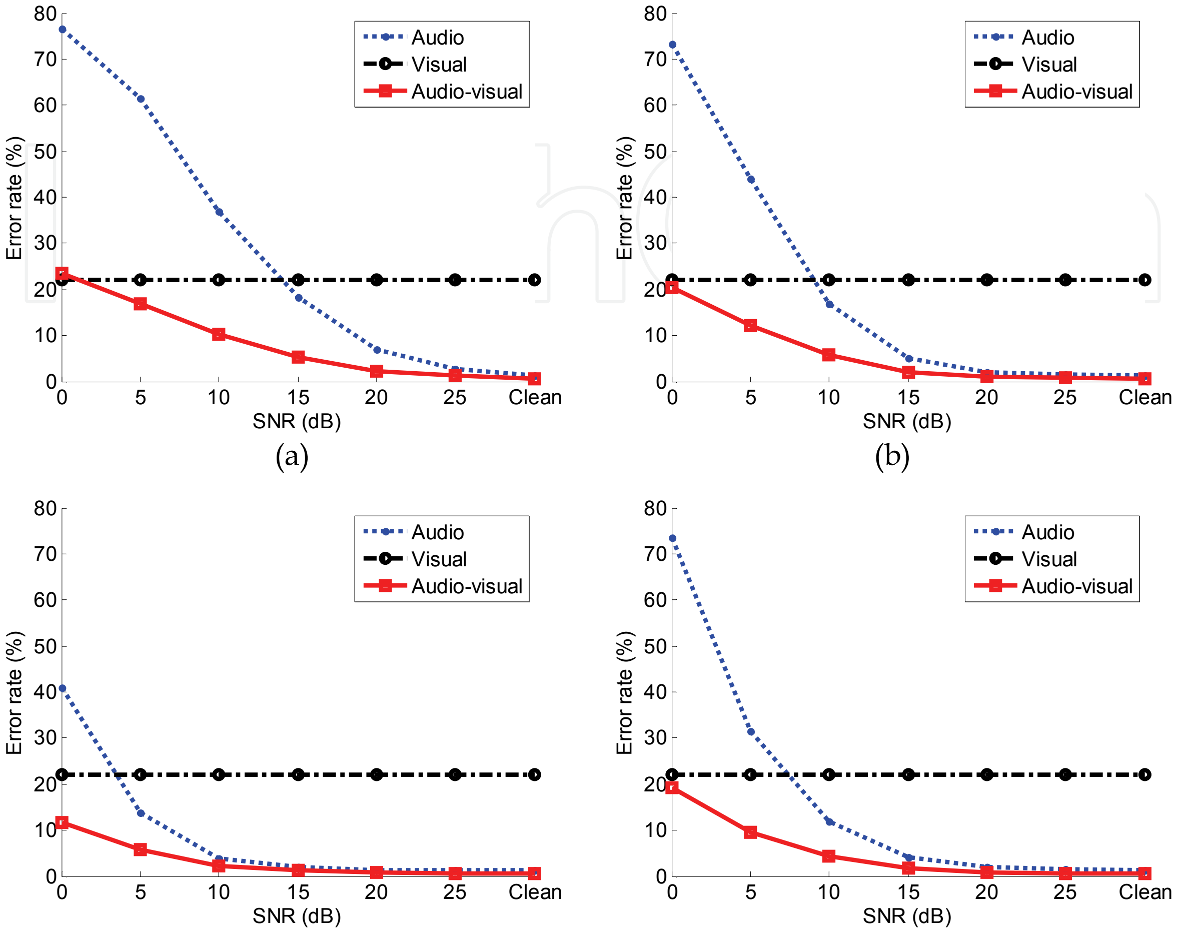
(b)

(c)

(d)

Fig. 12. Recognition performance of the unimodal and the bimodal systems in error rates (%) for the DIGIT database. (a) WHT. (b) F16. (c) FAC. (d) OPR.

Figure 14 shows the integration weight values (the means and the standard deviations) determined by the neural network with respect to SNRs for the DIGIT database. It is observed that the automatically determined weight value is large for high SNRs and small for low SNRs, as expected.



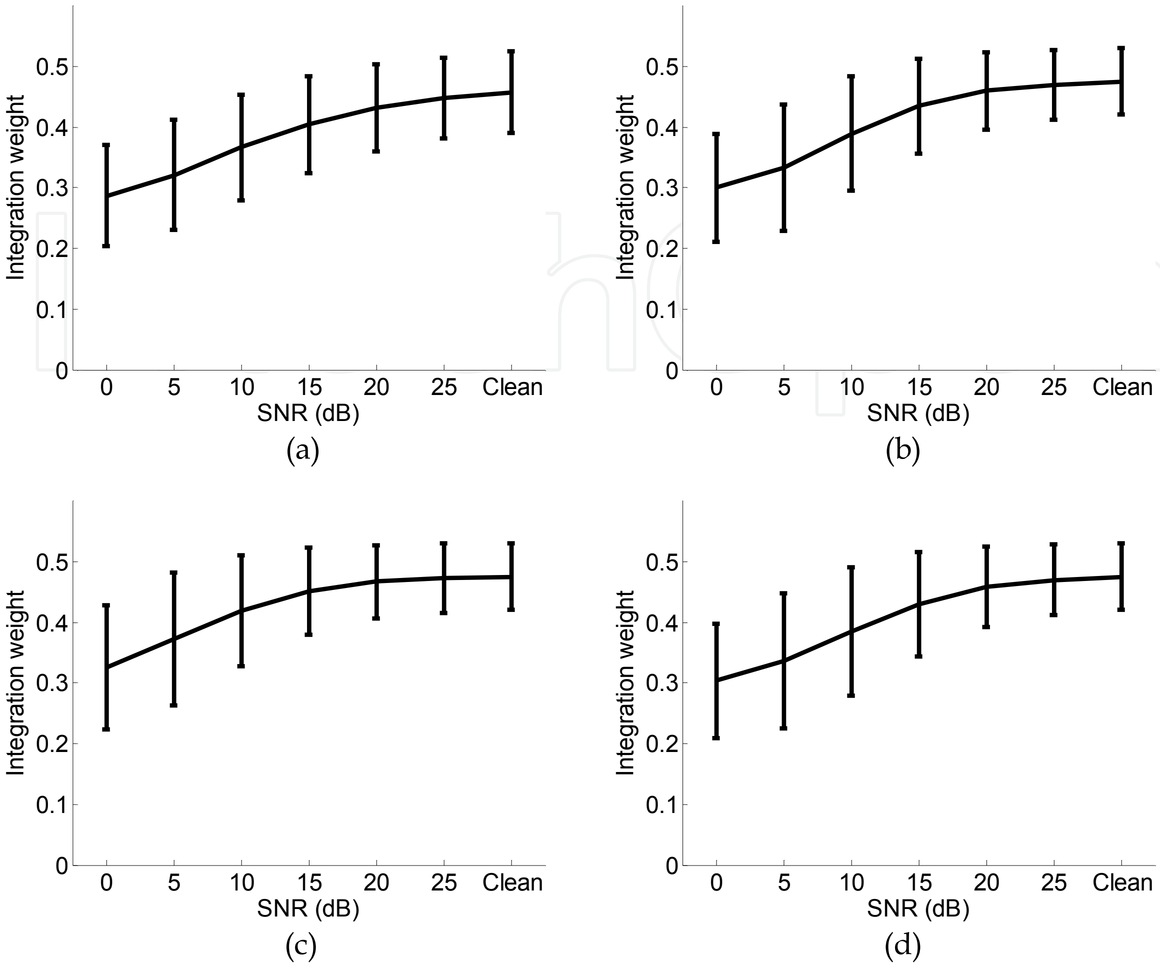
(a) (b) (c) (d)

Fig. 13. Recognition performance of the unimodal and the bimodal systems in error rates (%) for the CITY database. (a) WHT. (b) F16. (c) FAC. (d) OPR.

5. Conclusion

This chapter addressed the problem of information fusion for AVSR. We introduced the bimodal nature of speech production and perception by humans and defined the goal of audio-visual integration. We reviewed two existing approaches for implementing audio-visual fusion in AVSR systems and explained the preference of decision fusion to feature fusion for constructing noise-robust AVSR systems. For implementing a noise-robust AVSR system, different definitions of the reliability of a modality were discussed and compared. A neural network-based fusion method was described for effectively utilizing the reliability measures of the two modalities and producing noise-robust recognition performance over various noise conditions. It has been shown that we could successfully obtain the synergy of the two modalities.

The audio-visual information fusion method shown in this chapter mainly aims at obtaining robust speech recognition performance, which may lack modelling of complicated humans' audio-visual speech perception processes. If we consider that the humans' speech



(a)

(b)

(c)

(d)

Fig. 14. Generated integration weights with respect to the SNR value for the DIGIT database.

perception performance is surprisingly good, it is worth investigating such perception processes carefully and incorporating knowledge about them into implementing AVSR systems. Although it is still not clearly understood about such processes, it is believed that the two perceived signals complicatedly interact at multiple stages in humans' sensory systems and brains. As discussed in Section 2.1, visual information helps to detect acoustic speech under the presence of noise, which suggests that the two modalities can be used at the early stage of AVSR for speech enhancement and selective attention. Also, it has been suggested that there exist an early activation of auditory areas by visual cues and a later speech-specific activation of the left hemisphere possibly mediated by backward-projections from multisensory areas, which indicates that audio-visual interaction takes place in multiple stages sequentially (Hertrich et al., 2007). Further investigation of biological multimodal information processing mechanisms and modelling them for AVSR would be a valuable step toward mimicking humans' excellent AVSR performance.

6. References

Adjoudani, A. & Benoît, C. (1996). On the integration of auditory and visual parameters in an HMM-based ASR, In: *Speechreading by Humans and Machines: Models, Systems, and*

Applications, Stork, D. G. & Hennecke, M. E., (Eds.), pp. 461-472, Springer, Berlin, Germany.

Arnold, P. & Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, Vol. 92, (2001) pp. 339-355.

Benoît, C.; Mohamadi, T. & Kandel, S. D. (1994). Effects of phonetic context on audio-visual intelligibility of French. *Journal of Speech and Hearing Research*, Vol. 37, (October 1994) pp. 1195-1203.

Benoît, C. (2000). The intrinsic bimodality of speech communication and the synthesis of talking faces, In: *The Structure of Multimodal Dialogue II*, Taylor, M. M.; Nel, F. & Bouwhis, D. (Eds.), John Benjamins, Amsterdam, The Netherlands.

Braida, L. (1991). Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology Section A*, Vol. 43, No. 3, (August 1991) pp. 647-677.

Bregler, C. & Konig, Y. (1994). Eigenlips for robust speech recognition, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 669-672, Adelaide, Autralia, 1994.

Calvert, G. A.; Bullmore, E. T.; Brammer, M. J.; Campbell, R.; Williams, S. C. R.; McGuire, P. K.; Woodruff, P. W. R.; Iversen, S. D. & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, Vol. 276, (April 1997) pp. 593-596.

Campbell, R. (1988). Tracing lip movements: making speech visible. *Visible Language*, Vol. 22, No. 1, (1988) pp. 32-57.

Chibelushi, C. C.; Deravi, F. & Mason, J. S. D. (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, Vol. 4, No. 1, (March 2002) pp. 23-37.

Conrey, B. & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of Acoustical Society of America*, Vol. 119, No. 6, (June 2006) pp. 4065-4073.

Davis, S. B. & Mermelstein. (1980). Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, (1980) pp. 357-366.

Dupont, S. & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, Vol. 2, No. 3, (September 2000) pp. 141-151.

Ernest, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, Vol. 415, No. 6870, (January 2002) pp. 429-433.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, Vol. 14, (1986) pp. 3-28.

Gonzalez, R. C. & Woods, R. E. (2001). *Digital Image Processing*, Addison-Wesley Publishing Company.

Grant, K. W. & Seitz, P.-F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of Acoustical Society of America*, Vol. 103, No. 3, (September 2000) pp. 1197-1208.

Gurbuz, S.; Tufekci, Z.; Patterson, E. & Gowdy, J. (2001). Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 177-180, Salt Lake City, UT, USA, May 2001.

Hagan, M. T. & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, Vol. 5, No. 6, (1994) pp. 989-993.

Hertrich, I.; Mathiak, K.; Lutzenberger, W.; Menning, H. & Ackermann, H. (2007). Sequential audiovisual interactions during speech perception: a whole-head MEG study. *Neuropsychologia*, Vol. 45, (2007) pp. 1342-1354.

Hornik, K.; Stinchcombe, M. & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, Vol. 2, No. 5, (1989) pp. 359-366.

Huang, X.; Acero, A. & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice-Hall, Upper Saddle River, NJ, USA.

Kaynak, M. N.; Zhi, Q.; Cheok, A. D.; Sengupta, K.; Jian, Z. & Chung, K. C. (2004). Lip geometric features for human-computer interaction using bimodal speech recognition: comparison and analysis. *Speech Communication*, Vol. 43, No. 1-2, (January 2004) pp. 1-16.

Kim, J. & Davis, C. (2004). Investigating the audio-visual speech detection advantage. *Speech Communication*, Vol. 44, (2004) pp 19-30.

Klatt, D. H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, Vol. 7, (1979) pp. 279-312.

Lee, J.-S. & Park, C. H. (2006). Training hidden Markov models by hybrid simulated annealing for visual speech recognition, *Proceedings of the International Conference on Systems, Man and Cybernetics*, pp. 198-202, Taipei, Taiwan, October 2006.

Lee, J.-S. & Park, C. H. (2008). Robust audio-visual speech recognition based on late integration. *IEEE Transactions on Multimedia*, Vol. 10, No. 5, (August 2008) pp. 767-779.

Lewis, T. W. & Powers, D. M. W. (2004). Sensor fusion weighting measures in audio-visual speech recognition, *Proceedings of the Conference on Australasian Computer Science*, pp. 305-314, Dunedine, New Zealand, 2004.

Liberman, A. & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, Vol. 21, (1985) pp. 1-33.

Lucey, S. (2003). An evaluation of visual speech features for the tasks of speech and speaker recognition, *Proceedings of International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 260-267, Guilford, UK, June 2003.

Macaluso, E.; George, N.; Dolan, R; Spence, C. & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. *NeuroImage*, Vol. 21, (2004) pp. 725-732.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*, Erlbaum.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press.

Massaro, D. W. (1999). Speechreading: illusion or window into pattern recognition. *Trends in Cognitive Sciences*, Vol. 3, No. 8, (August 1999) pp. 310-317.

Massaro, D. W. & Stork, D. G. (1998). 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, No. 3, (May-June 1998) pp. 236-242.

Matthews, I.; Bangham, J. A. & Cox. S. (1996). Audio-visual speech recognition using multiscale nonlinear image decomposition, *Proceedings of the International Conference on Speech and Language Processing*, pp. 38-41, Philadelphia, USA, 1996.

Matthews, I.; Potamianos, G.; Neti, C. & Luettin, J. (2001). A comparison of model and transform-based visual features for audio-visual LVCSR, *Proceedings of the International Conference on Multimedia and Expo*, pp. 22-25, Tokyo, Japan, April 2001.

McClelland, J. L. & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, Vol. 18, (1986) pp. 1-86.

McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, Vol. 264, (December 1976) pp. 746-748.

Pekkola, J.; Ojanen, V.; Autti, T.; Jääskeläinen. I. P.; Möttönen, R.; Tarkiainen, A. & Sams, M. (2005). Primary auditory cortex activation by visual speech: an fMRI study at 3T. *NeuroReport*, Vol. 16, No. 2, (February 2005) pp. 125-128.

Potamianos, G. & Neti, C. (2000). Stream confidence estimation for audio-visual speech recognition, *Proceedings of the International Conference on Spoken Language Processing*, pp. 746-749, Beijing, China, 2000.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, (Febrary 1989) pp. 257-286.

Robert-Ribes, J.; Schwartz, J.-L.; Lallouache, T. & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of Acoustical Society of America*, Vol. 103, No. 6, (June 1998) pp. 3677-3689.

Rogozan, A & Deléglise, P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, Vol. 26, No. 1-2, (October 1998) pp. 149-161.

Ross, L. A.; Saint-Amour, D.; Leavitt, V. M.; Javitt, D. C. & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, Vol. 17, No. 5, (May 2007) pp. 1147-1153.

Ruytjens, L.; Albers, F.; van Dijk, P.; Wit, H. & Willemsen, A. (2007). Activation in primary auditory cortex during silent lipreading is determined by sex. *Audiology and Neurotology*, Vol. 12, (2007) pp. 371-377.

Schwartz, J.-L.; Berthommier, F. & Savariaux, C. (2004). Seeing to hear better : evidence for early audio-visual interactions in speech identification. *Cognition*, Vol. 93, (2004) pp. B69-B78.

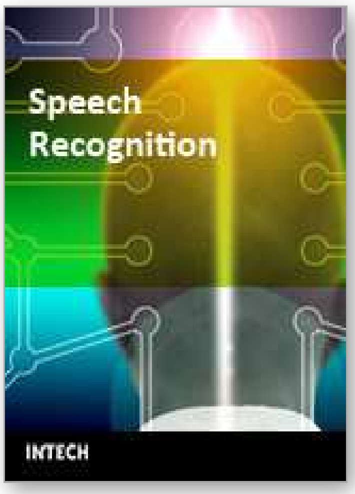
Sekiyama, K. & Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, Vol. 21, (1993) pp. 427-444.

Sharma, R.; Pavlović, V. I. & Huang, T. S. (1998). Toward multimodal human-computer interface. *Proceedings of the IEEE*, Vol. 86, No. 5, (May 1998) pp. 853-869.

Stein, B. & Meredith, M. A. (1993). *The Merging of Senses*, MIT Press, MA, USA.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception, In: *Hearing by Eye: The Psychology of Lip-reading*, Dodd, B. & Campbell, R. (Eds.), pp. 3-51, Lawrence Erlbarum, London, UK.

Varga, A. & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, Vol. 12, No. 3, (1993) pp. 247-251.



Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9
Hard cover, 550 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jong-Seok Lee and Cheol Hoon Park (2008). Adaptive Decision Fusion for Audio-Visual Speech Recognition, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/adaptive_decision_fusion_for_audio-visual_speech_recognition

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821