15

Towards Augmentative Speech Communication

Panikos Heracleous[1], Denis Beautemps[2], Hiroshi Ishiguro[3], and Norihiro Hagita[4]

[1,3,4]*ATR, Intelligent Robotics and Communication Laboratories Japan Science and Technology Agency, CREST*

[2]*GIPSA-lab, Speech and Cognition Department, CNRS-Grenoble University*

[1,3,4]*Japan,*

[2]*France*

1. Introduction

Speech is the most natural form of communication for human beings and is often described as a unimodal communication channel. However, it is well known that speech is multimodal in nature and includes the auditive, visual, and tactile modalities. Other less natural modalities such as electromyographic signal, invisible articulator display, or brain electrical activity or electromagnetic activity can also be considered. Therefore, in situations where audio speech is not available or is corrupted because of disability or adverse environmental condition, people may resort to alternative methods such as augmented speech.

In several automatic speech recognition systems, visual information from lips/mouth and facial movements has been used in combination with audio signals. In such cases, visual information is used to complement the audio information to improve the system's robustness against acoustic noise (Potamianos et al., 2003).

For the orally educated deaf or hearing-impaired people, lip reading remains a crucial speech modality, though it is not sufficient to achieve full communication. Therefore, in 1967, Cornett developed the Cued Speech system as a supplement to lip reading (O.Cornett, 1967). Recently, studies have been presented on automatic Cued Speech recognition using hand gestures in combination with lip/mouth information (Heracleous et al., 2009).

Several other studies have been introduced that deal with the problem of alternative speech communication based on speech modalities other than audio speech. A method for communication based on inaudible speech received through body tissues has been introduced using the Non-Audible Murmur (NAM) microphone. NAM microphones have been used for receiving and automatically recognizing sounds of speech-impaired people, for ensuring privacy in communication, and for achieving robustness against noise (Heracleous et al., 2007; Nakamura et al., 2008). Aside from automatic recognition of NAM speech, silicon NAM microphones were used for NAM-to-speech conversion (Toda & Shikano, 2005; Tran et al., 2008).

A few researchers have addressed the problem of augmented speech based on the activation signal of the muscles produced during speech production (Jou et al., 2006). The OUISPER project (Hueber et al., 2008) attempts to automatically recognize and resynthesize speech based on the signals of tongue movements captured by an ultrasound device in combination with lip information.
In this article, automatic recognition of Cued Speech for French and Non-Audible Murmur (NAM) recognition are introduced. Cued Speech is a visual mode for communication in the deaf society. Using only visual information produced by lip movements and hand shapes, all the sounds of a spoken language can be visually distinguished and thus enabling deaf individuals to communicate with each other and also with normal-hearing people. Non-Audible Murmur is very quietly uttered speech which can be perceived by a special acoustic sensor (i.e., NAM microphone). NAM microphones can be used for privacy, for robustness against noise, and also by speech-impaired people. In this study, experimental results are also presented showing the effectiveness of the two methods in augmentative speech communication.

2. Cued Speech

To date, visual information is widely used to improve speech perception or automatic speech recognition (lipreading) (Potamianos et al., 2003). With lipreading technique, speech can be understood by interpreting the movements of lips, face and tongue. In spoken languages, a particular facial and lip shape corresponds to a specific sound (phoneme). However, this relationship is not one-to-one and many phonemes share the same facial and lip shape (visemes). It is impossible, therefore to distinguish phonemes using visual information alone. Without knowing the semantic context, one cannot perceive the speech thoroughly even with high lipreading performances. To date, the best lip readers are far away into reaching perfection. On average, only 40 to 60% of the vowels of a given language (American English) are recognized by lipreading (Montgomery & Jackson, 1983), and 32% when relating to low predicted words (Nicholls & Ling, 1982). The best result obtained amongst deaf participants was 43.6% for the average accuracy (Auer & Bernstein, 2007; Bernstein et al., 2007). The main reason for this lies in the ambiguity of the visual pattern. However, as far as the orally educated deaf people are concerned, the act of lipreading remains the main modality of perceiving speech.
To overcome the problems of lipreading and to improve the reading abilities of profoundly deaf children, Cornett (O.Cornett, 1967) developed in 1967 the Cued Speech system to complement the lip information and make all phonemes of a spoken language clearly visible. As many sounds look identical on face/lips (e.g., /*p*/, /*b*/, and /*m*/), using hand information those sounds can be distinguished and thus make possible for deaf people to completely understand a spoken language using visual information only.
Cued Speech [also referred to as Cued Language (Fleetwood & Metzger, 1998)] uses hand shapes placed in different positions near the face along with natural speech lipreading to enhance speech perception from visual input. This is a system where the speaker faces the perceiver and moves his hand in close relation with speech. The hand, held flat and oriented so that the back of the hand faces the perceiver, is a cue that corresponds to a unique phoneme when associated with a particular lip shape. A manual cue in this system contains two components: the hand shape and the hand position relative to the face. Hand shapes distinguish among consonant phonemes whereas hand positions distinguish among vowel

phonemes. A hand shape, together with a hand position, cues a syllable. Cued Speech

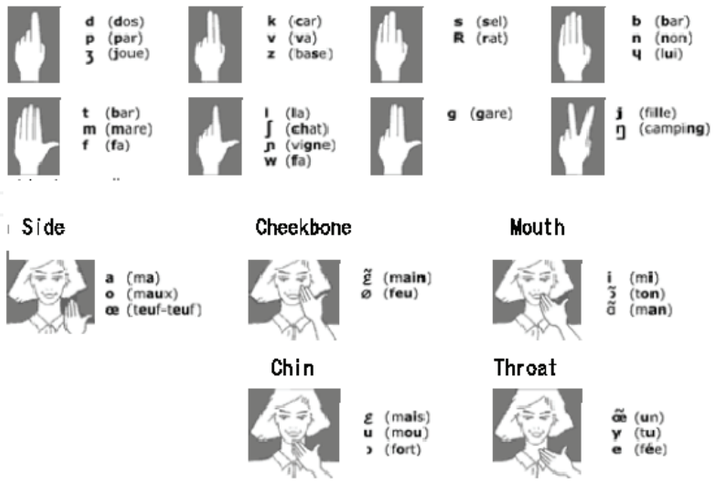

Fig. 1. Hand shapes for consonants (top) and hand position (bottom) for vowels in French Cued Speech.

improves the speech perception of deaf people (Nicholls & Ling, 1982; Uchanski et al., 1994). Moreover, for deaf people who have been exposed to this mode since their youth, it offers a complete representation of the phonological system, and therefore it has a positive impact on the language development (Leybaert, 2000). Figure 1 describes the complete system for French. In French Cued Speech, eight hand shapes in five positions are used. The system was adapted from American English to French in 1977. To date, Cued Speech has been adapted in more than 60 languages.

Another widely used communication method for deaf individuals is the Sign Language (Dreuw et al., 2007; Ong & Ranganath, 2005). Sign Language is a language with its own grammar, syntax and community; however, one must be exposed to native and/or fluent users of Sign Language to acquire it. Since the majority of children who are deaf or hard-of-hearing have hearing parents (90%), these children usually have limited access to appropriate Sign Language models. Cued Speech is a visual representation of a spoken language, and it was developed to help raise the literacy levels of deaf individuals. Cued Speech was not developed to replace Sign Language. In fact, Sign Language will be always a part of deaf community. On the other hand, Cued Speech is an alternative communication method for deaf individuals. By cueing, children who are deaf would have a way to easily acquire the native home language, read and write proficiently, and communicate more easily with hearing family members who cue them.

In the first attempt for vowel recognition in Cued Speech, in (Aboutabit et al., 2007) a method based on separate identification, i.e., indirect decision fusion was used and a 77.6% vowel accuracy was obtained. In this study, however, the proposed method is based on HMMs and uses concatenative feature fusion to integrate the components into a combined one and then perform automatic recognition. Fusion (Adjoudani & Benoît, 1996; Hennecke et al., 1996; Nefian et al., 2002) is the integration of all available single modality streams into a combined one. In this study, lip shape and hand components are combined in order to realize automatic recognition in Cued Speech for French.

3. Non-Audible Murmur (NAM)

Non-Audible Murmur (NAM) refers to a very softly uttered speech received through the body tissue. A special acoustic sensor (i.e., the NAM microphone) is attached behind the talker's ear. This receives very soft sounds that are inaudible to other listeners who are in close proximity to the talker.

The attachment of the NAM microphone to the talker is shown in Figure 2. The first NAM microphone was based on stethoscopes used by medical doctors to examine patients, and was called the stethoscopic microphone (Nakajima et al., 2003). Stethoscopic microphones were used for the automatic recognition of NAM speech (Heracleous et al., 2004). The silicon NAM microphone is a more advanced version of the NAM microphone (Nakajima et al., 2005). The silicon NAM microphone is a highly sensitive microphone wrapped in silicon; silicon is used because its impedance is similar to that of human skin. Silicon NAM microphones have been employed for automatic recognition of NAM speech as well as for NAM-to-speech conversion (Toda & Shikano, 2005). Similar approaches have been introduced for speech enhancement or speech recognition (Jou et al., 2004; Zheng et al., 2003). Further, non-audible speech recognition has also been reported based on electromyographic (EMG) speech recognition, which processes electric signals caused by the articulatory muscles (Walliczek et al., 2006).

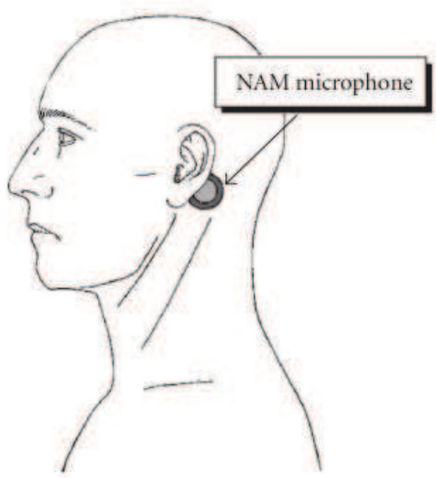

Fig. 2. NAM microphone attached to the talker

The speech received by a NAM microphone has different spectral characteristics in comparison to normal speech. In particular, the NAM speech shows limited high-frequency contents because of body transmission. Frequency components above the 3500-4000 Hz range are not included in NAM speech. The NAM microphone can also be used to receive audible speech directly from the body [Body Transmitted Ordinary Speech (BTOS)]. This enables automatic speech recognition in a conventional way while taking advantage of the robustness of NAM against noise.

Previous studies have reported experiments for NAM speech recognition that produced very promising results. A word accuracy of 93.9% was achieved for a 20k Japanese vocabulary dictation task when a small amount of training data from a single speaker was used (Heracleous et al., 2004). Moreover, experiments were conducted using simulated and real

noisy test data with clean training models to investigate the role of the Lombard reflex (Heracleous et al., 2007; Junqua, 1993) in NAM recognition.
In the present study, audio-visual NAM recognition is investigated by using the concatenative feature fusion, the multistream HMM decision fusion, and late fusion to integrate the audio and visual information. A statistical significance test was performed, and audio-visual NAM recognition in a noisy environment was also investigated.

4. Experiments

4.1 Cued Speech automatic recognition

The data for vowel- and consonant recognition experiments were collected from a normal-hearing cuer. The female native French speaker employed for data recording was certified in transliteration speech into Cued Speech in the French language. She regularly cues in schools. The cuer wore a helmet to keep her head in a fixed position and opaque glasses to protect her eyes against glare from the halogen floodlight. The cuer's lips were painted blue, and blue marks were marked on her glasses as reference points. These constraints were applied in recordings in order to control the data and facilitate the extraction of accurate features.
The data were derived from a video recording of the cuer pronouncing and coding in Cued Speech a set of 262 French sentences. The sentences (composed of low predicted multi-syllabic words) were derived from a corpus that was dedicated to Cued Speech synthesis (Gibert et al., 2005). Each sentence was dictated by an experimenter, and was repeated two or three times (to correct the pronunciation errors) by the cuer resulting in a set of 638 sentences.

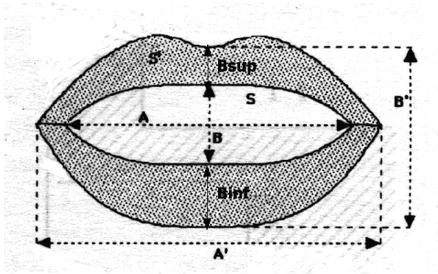

Fig. 3. Parameters used for lip shape modeling.

The audio part of the video recording was synchronized with the image. Figure 3 shows the lip shape parameters used in the study. An automatic image processing method was applied to the video frames in the lip region to extract their inner and outer contours and derive the corresponding characteristic parameters: lip width (A), lip aperture (B), and lip area (S) (i.e., six parameters in all).
The process described here resulted in a set of temporally coherent signals: the 2D hand information, the lip width (A), the lip aperture (B), and the lip area (S) values for both inner and outer contours, and the corresponding acoustic signal. In addition, two supplementary parameters relative to the lip morphology were extracted: the pinching of the upper lip (Bsup) and lower (Binf) lip. As a result, a set of eight parameters in all was extracted for modeling lip shapes. For hand position modeling, the xy coordinates of two landmarks placed on the hand were used (i.e., 4 parameters). For hand shape modeling, the xy coordinates of the landmarks

placed on the fingers were used (i.e., 10 parameters). Non-visible landmarks received default coordinates [0,0].

During the recording of Cued Speech material for isolated word recognition experiments, the conditions were different from the ones described earlier. The system was improved by excluding the use of a helmet by the cuer, enabling in this way the head movements during recording. The subject was seated on a chair in a way to avoid large movements in the third direction (i.e. towards the camera). However, the errors that might occur have not been evaluated. In addition, the landmarks placed on the cuer's fingers were of different colors in order to avoid the hand shape coding and the finger identification, and this helped to simplify and speed up the image processing stage. In these recording sessions, a normal-hearing cuer and a deaf cuer were employed. The corpus consisted of 1450 isolated words with each of 50 words repeated 29 times by the cuers.

In the phoneme recognition experiments, context-independent, 3-state, left-to-right, no-skip-phoneme HMMs were used. Each state was modeled with a mixture of 32 Gaussians. In addition to the basic lip and hand parameters, first- (Δ) and second-order derivatives ($\Delta\Delta$) were used as well. For training and test, 426 and 212 sentences were used, respectively. The training sentences contained 3838 vowel and 4401 consonant instances, and the test sentences contained 1913 vowel and 2155 consonant instances, respectively. Vowels and consonants were extracted automatically from the data after a forced alignment was performed using the audio signal.

For isolated word recognition experiments two HMM sets were trained (deaf and normal-hearing). Fifteen repetitions of each word were used to train 50, 6-state, whole word HMMs, and 14 repetitions were used for testing. Eight and ten parameters were used for lip shape and hand shape modeling, respectively.

In automatic speech recognition, a diagonal covariance matrix is often used because of the assumption that the parameters are uncorrelated. In lipreading, however, parameters show a strong correlation. In this study, a global Principal Component Analysis (PCA) using all the training data was applied to decorrelate the lip shape parameters and then a diagonal covariance matrix was used. The test data were then projected into the PCA space. All PCA lip shape components were used for HMM training. For training and recognition the HTK3.1 toolkit (Young et al., 2001) was used.

For the integration of the lip shape and hand shape components, feature concatenative fusion was used. Feature concatenation uses the concatenation of the synchronous lip shape and hand features as the joint feature vector

$$O_t^{LH} = [O_t^{(L)^T}, O_t^{(H)^T}]^T \in R^D \quad (1)$$

where O_t^{LH} is the joint lip-hand feature vector, $O_t^{(L)}$ the lip shape feature vector, $O_t^{(H)}$ the hand feature vector, and D the dimensionality of the joint feature vector. In vowel recognition experiments, the dimension of the lip shape stream was 24 (8 basic parameters, 8 Δ, and 8 $\Delta\Delta$ parameters) and the dimension of the hand position stream was 12 (4 basic parameters, 4 Δ, and 4 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand position feature vectors was, therefore 36. In consonant recognition experiments, the dimension of the hand shape stream was 30 (10 basic parameters, 10 Δ, and 10 $\Delta\Delta$ parameters). The dimension D of the joint lip-hand shape feature vectors was, therefore 54. Figure 4 shows the vowel recognition results. As shown, by integrating hand position component with lip shape component, a

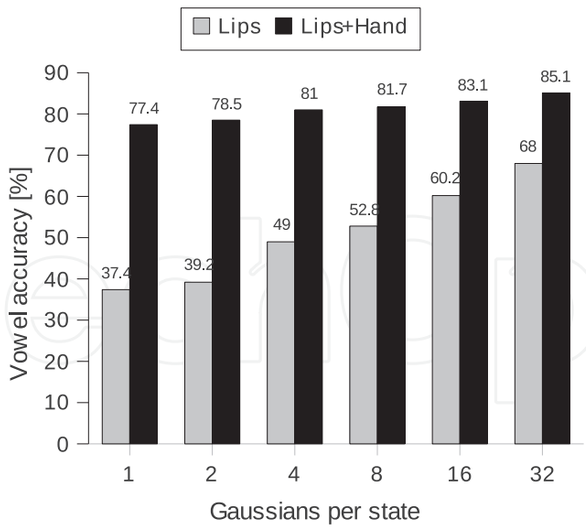

Fig. 4. Cued Speech vowel recognition using only lip and hand parameters based on concatenative feature fusion.

vowel accuracy of 85.1% was achieved, showing a 53% relative improvement compared to the sole use of lip shape parameters.

Using concatenative feature fusion, lip shape component was integrated with hand shape component and consonant recognition was conducted. For hand shape modeling, the *xy* coordinates of the fingers, and first- and second-order derivatives were used. In total, 30 parameters were used for hand shape modeling. For lip shape modeling, 24 parameters were used. Figure 5 shows the obtained results in the function of Gaussians per state. It can be seen that when using 32 Gaussians per state, a consonant accuracy of 78.9% was achieved. Compared to the sole use of lip shape, a 56% relative improvement was obtained.

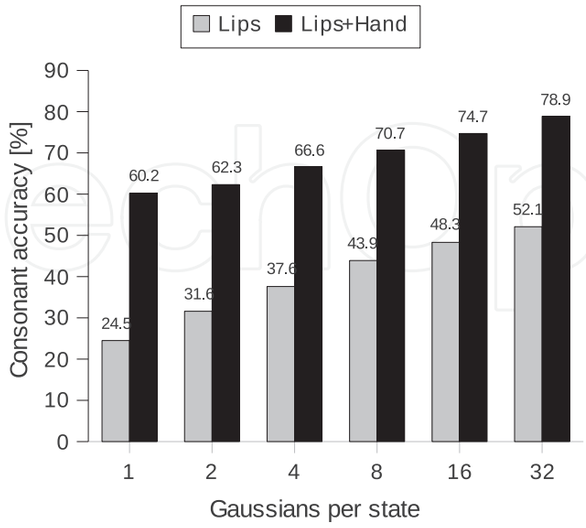

Fig. 5. Cued Speech consonant recognition using only lip and hand parameters based on concatenative feature fusion.

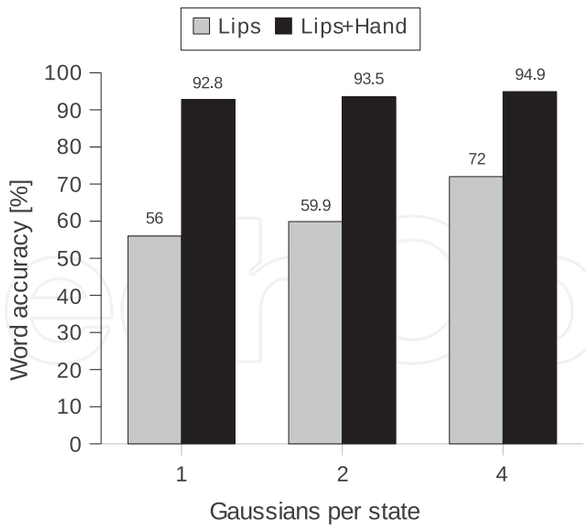

Fig. 6. Word accuracy for isolated word recognition in the case of a normal-hearing subject.

Figure 6 shows the isolated word recognition results obtained in the function of several Gaussians per state in the case of the normal-hearing cuer. In the case of a single Gaussian per state, using lip shape alone obtained a 56% word accuracy; however, when hand shape information was also used, a 92.8% word accuracy was obtained. The highest word accuracy when using lip shape was 72%, obtained in the case of using 4 Gaussians per state. In that case, the Cued Speech word accuracy using also hand information was 94.9%. Figure 7 shows the

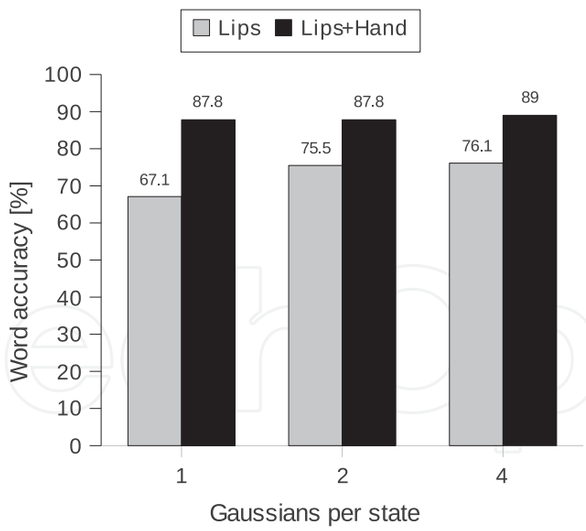

Fig. 7. Word accuracy for isolated word recognition in the case of a deaf subject.

obtained results in the case of a deaf cuer. The results show that in the case of the deaf subject, words were better recognized when using lip shape alone compared to the normal-hearing subject. The fact that deafs rely on lipreading for speech communication may increase their ability not only for speech perception but also for speech production. The word accuracy in the case of the deaf subject was 89% compared to the 94.9% in the normal-hearing subject.

	HMMs		
Test data	Normal	Deaf	Normal+Deaf
Normal	94.9	0.6	92.0
Deaf	2.0	89.0	87.2

Table 1. Word accuracy of a multi-speaker experiment

The difference in performance might be because of the lower hand shape recognition in the deaf subject. It should also be noted that the normal-hearing cuer was a professional teacher of Cued Speech. The results show that there are no additional difficulties in recognizing Cued Speech in deaf subjects, other than those appearing in normal-hearing subjects.
A multi-cuer isolated word recognition experiment was also conducted using the normal-hearing and the deaf cuers' data. The aim of this experiment is to investigate whether it is possible to train speaker-independent HMMs for Cued Speech recognition. The training data consisted of 750 words from the normal-hearing subject, and 750 words from the deaf subject. For testing, 700 words from normal-hearing subject and 700 words from the deaf subject were used, respectively. Each state was modeled with a mixture of 4 Gaussian distributions. For lip shape and hand shape integration, the concatenative feature fusion was used.
Table 1 shows the results obtained when lip shape and hand shape features were used. The results show, that due to the large variability between the two subjects, word accuracy of cross-recognition is extremely low. On the other hand, the word accuracy in normal-hearing subject when using multi-speaker HMMs was 92%, which is comparable with the 94.9% word accuracy when cuer-dependent HMMs were used. In the case of the deaf subject, the word accuracy when using multi-cuer HMMs was 87.2%, which was also comparable with the 89% word accuracy when using speaker-dependent HMMs.
The results obtained indicate that creating speaker-independent HMMs for Cued Speech recognition using a large number of subjects should not face any particular difference, other than those appear in the conventional audio speech recognition. To prove this, however, additional experiments using a large number of subjects are required.

4.2 NAM automatic recognition

The corpus used in the experiment was 212 continuous Japanese utterances, containing 7518 phoneme realisations. A 3-state with no skip HMM topology was used. Forty-three monophones were trained using 5132 phonemes. For the purpose of testing, 2386 phonemes were used. The audio parameter vectors were of length 36 (12 MFCC, 12ΔMFCC, and 12 ΔΔMFCC). The HTK3.4 Toolkit was used for training and testing.
The face and profile views of the subject were filmed under conditions of good lighting. The system captured the 3-D positions of 112 colored beads glued on the speaker's face at a sampling rate of 50 Hz (fig. 8), synchronized with the acoustic signal sampled at 16000 Hz. The collection of 30 lip points using a generic 3-D geometric model of the lips is shown in Figure 9 (Revéret & Benoît, 1998).
The shape model is built using the Principal Component Analysis (PCA). Successive applications of PCA are performed on the selected subsets of the data, which generate the main directions. These directions are retained as linear predictors for the whole data set. The mobile points P of the face deviate from their average position B by a linear composition of

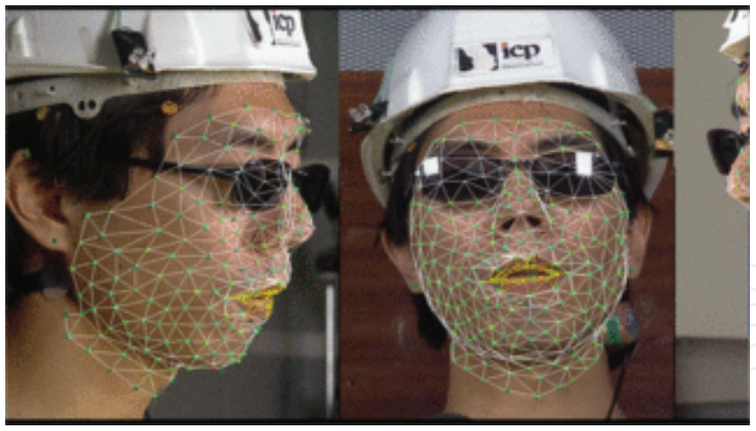

Fig. 8. Characteristic points used for capturing the movements.

the basic components M loaded by factors α (articulatory parameters) (Revéret et al., 2000).

$$P = B + \alpha M \tag{2}$$

Only the first 5 parameters of the extracted 12 linear components M were used. These explained more than 90% of the data variance using the following iterative linear prediction on the data residual: the first component of the PCA on the lower teeth (LT) values leads to the "first jaw" predictor. The PCA on the residual lips values (without jaw1 influence) usually presented three pertinent lip predictors (i.e., lips protrusion, lips closing mainly required for bilabials, and lips raising mainly required for labiodental fricatives). The movements of the throat linked the underlying movements of the larynx and the hyoid bone, and served as the fifth one. The video parameters were interpolated at 200 Hz to synchronize with the audio analysis frame rate. For audio-visual NAM recognition, concatenative feature fusion,

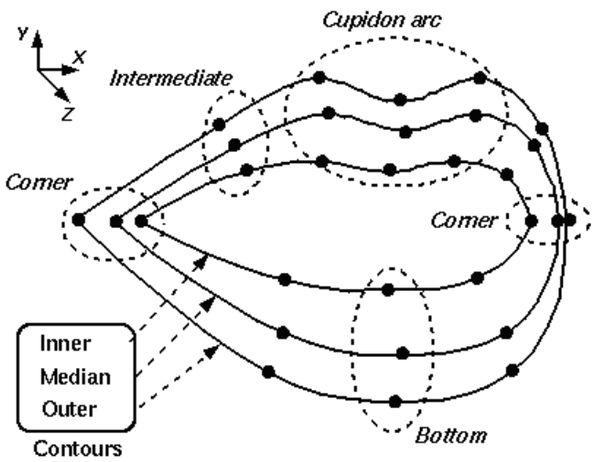

Fig. 9. The 30 control points and the 3 basic contour curves.

multistream decision fusion, and late fusion methods were used.
Multistream HMM fusion is a state synchronous decision fusion, which captures the reliability of each stream by combining the likelihoods of single-stream HMM classifiers (Potamianos

et al., 2003). The emission likelihood of the multistream HMM is the product of the emission likelihoods of the single-stream components, weighted appropriately by stream weights. Given the O combined observation vector, that is, the NAM and visual elements, the emission score of multistream HMM is given by:

$$b_j(O_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm})]^{\lambda_s} \tag{3}$$

where, $N(O; \mu, \Sigma)$ is the value in O of a multivariate Gaussian with mean μ and covariance matrix Σ, and S is the number of the streams. For each stream s, M_s Gaussians in a mixture are used, each weighted with c_{jsm}. The contribution of each stream is weighted by λ_s. In the present study, it is assumed that the stream weights do not depend on state j and time t. However, two constraints were applied, namely:

$$0 \leq \{\lambda_n, \lambda_v\} \leq 1, \quad and \quad \lambda_n + \lambda_v = 1 \tag{4}$$

where λ_n is the NAM stream weight, and λ_v is the visual stream weight. In these experiments, the weights were experimentally adjusted to 0.6 and 0.4 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

A disadvantage of the previously described fusion methods is the assumption that there is a synchrony between the two streams. In the present study, late fusion was applied to enable asynchrony between the NAM stream and the visual stream. In the late fusion method, two single HMM-based classifiers were used for the NAM speech and the visual speech, respectively. For each test utterance (i.e., isolated phone), the two classifiers provided an output list, which included all the phone hypotheses with their likelihoods. Subsequently, all the separate mono-modal hypotheses were combined into the bi-modal hypotheses using the weighted likelihoods, as given by:

$$logP_{NV}(h) = \lambda_n logP_N(h|\underline{O}_N) + \lambda_v logP_V(h|\underline{O}_V) \tag{5}$$

where, $logP_{NV}(h)$ is the score of the combined bi-modal hypothesis h, $logP_N(h|\underline{O}_N)$ is the score of the h provided by the NAM classifier, and $logP_V(h|\underline{O}_V)$ is the score of the h provided by the visual classifier. λ_n and λ_v are the stream weights with the same constraints applied in multi-stream HMM fusion.

The procedure described in this study finally resulted in a combined N-best list in which the top hypothesis was selected as the correct bi-modal output. A similar method was also introduced in (Potamianos et al., 2003).

A comparison of the three classification methods used in the present study is shown in Table 2. As seen in the table, the highest classification accuracies are achieved when late fusion is used. The second best classification accuracies are achieved when using multistream HMM decision

	Fusion Method		
	Late	Multistream	Feature
Phonemes	71.8	68.9	67.8
Vowels	86.2	83.7	83.3
Consonants	64.1	59.7	58.2

Table 2. Comparison of the fusion methods in NAM automatic recognition.

fusion. Finally, the lowest accuracies are observed when using feature fusion. Specifically, when using late fusion, an accuracy of 71.8% is achieved for phoneme classification, 86.2% accuracy for vowel classification, and 64.1% accuracy for consonant classification. The highest accuracies, when using late fusion, might be an evidence of asynchrony between the NAM speech and the visual stream. In the following experiments late fusion is used to integrate the NAM audio speech with the visual data. The results obtained when using visual data, NAM

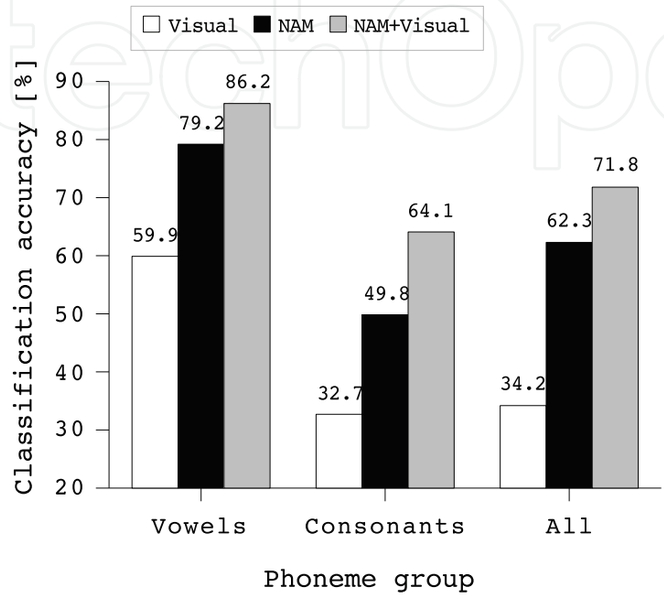

Fig. 10. Phoneme classification in a clean environment.

data, and visual-NAM data are shown in Figure 10. The results indicate that the classification accuracy is very low when only visual data is used. As many sounds appear to be similar on the lips/face, the sole use of visual parameters cannot distinguish these sounds. In the case of NAM data, the accuracies are higher in comparison to visual data. Specifically, an accuracy of 79.2% was achieved for vowel recognition, 49.8% accuracy for consonant recognition, and 59.7% accuracy for phoneme recognition. It is observed that the accuracy is considerably lower for consonant recognition in comparison to vowel recognition. However, because of the unvoiced nature of NAM, both voiced and unvoiced sounds articulated at the same place become similar, resulting in a larger number of confusions between consonants.The significant improvements in accuracy, when visual data were fused with NAM speech, are shown in Figure 10. Specifically, a relative improvement of 33% was achieved for vowel recognition, 28% for consonant recognition, and 30% for phoneme recognition.

The McNemar's test (Gillick & Cox, 1989) was performed to determine whether the differences were statistically significant . The p-values in all the cases were 0.001, which indicated that the differences were statistically significant.

Table 3 and Table 4 show the confusion matrices of the plosives sounds when using NAM and NAM-visual speech, respectively. As is shown, the number of confusions decreases when visual information was also used resulting in a higher accuracy.

In another experiment, office noise recorded by a NAM microphone was superimposed on clean NAM speech on several Signal-to-Noise-Ratio (SNR) levels. The noisy data were used

	/p/	/b/	/t/	/d/	/k/	/g/
/p/	0	0	5	0	2	1
/b/	0	8	5	1	3	0
/t/	2	0	36	3	12	1
/d/	0	2	6	14	4	1
/k/	1	0	8	0	45	6
/g/	0	1	0	2	6	20

Table 3. Confusion matrix of Japanese plosives using NAM speech.

	/p/	/b/	/t/	/d/	/k/	/g/
/p/	3	0	5	0	0	0
/b/	1	13	3	0	0	0
/t/	0	0	39	1	13	1
/d/	0	0	7	17	1	2
/k/	0	0	7	0	50	3
/g/	0	0	0	1	6	22

Table 4. Confusion matrix of Japanese plosives using NAM-visual speech.

to train HMMs of a desired SNR level. In addition, the noisy NAM data were fused with the visual parameters and audiovisual NAM HMMs were trained.
The classification accuracies in the function SNR levels for the visual, the NAM, and the NAM-visual cases are shown in Figure 11 . As seen in the figure, the accuracy of NAM recognition decreases when noisy data is used. However, the accuracy drastically increases when NAM speech is integrated with visual information. In such a case, an average of 15% absolute increase in accuracy was obtained.

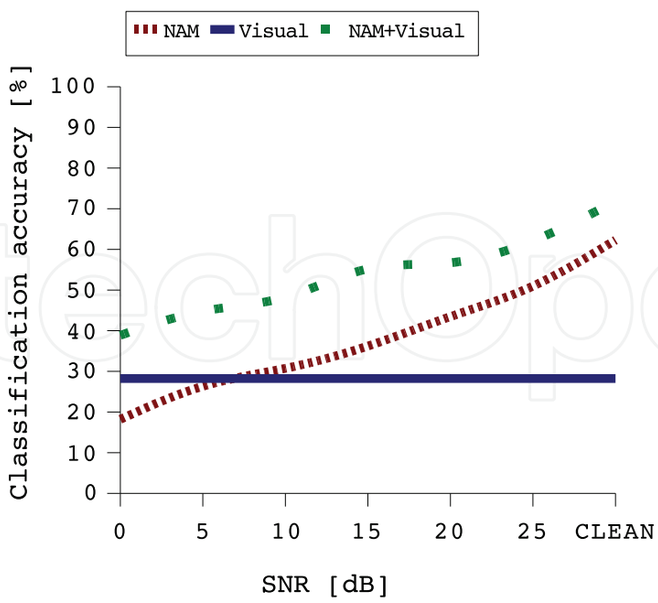

Fig. 11. Phoneme classification in noisy environment.

5. Conclusion and future word

In this chapter, two methods for augmentative speech communication were introduced. Specifically, automatic recognition for Cued Speech for French and Non-Audible Murmur recognition were reported. The authors demonstrated the effectiveness of both methods in alternative speech communication, when modalities other than the audio one are used. Regarding Cued Speech automatic recognition, the experimental results obtained showed recognition rates comparable to those obtained when audio speech is used. In addition, the results showed that using hand information as complement to lip movements, significantly higher rates achieved compared to the sole use of lip movements. With concern to Non-Audible Murmur recognition, the results showed that the unvoiced nature of NAM speech causes a higher number of confusions. Using, however, visual information produced by face/lips further improvements achieved compared with using NAM speech only. As future work, the authors plan to investigate the Cued Speech for Japanese, and also to evaluate the intelligibility of audible NAM speech in clean and noisy environments. This work has been partially supported by JST CREST 'Studies on Cellphone-type Teleoperated Androids Transmitting Human Presence'

6. References

Aboutabit, N., Beautemps, D. & Besacier, L. (2007). Automatic identification of vowels in the cued speech context, *in Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP)* .

Adjoudani, A. & Benoît, C. (1996) On the integration of auditory and visual parameters in an hmm-based asr, *in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany:Springer* p. 461471.

Auer, E. T. & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment, *Journal of Speech, Language, and Hearing* 50: 1157–1165.

Bernstein, L., Auer, E. & Jiang, J. (2007). Lipreading, the lexicon, and cued speech, *In C. la Sasso and K. Crain and J. Leybaert (Eds.), Cued Speech and Cued Language for Children who are Deaf or Hard of Hearing, Los Angeles, CA: Plural Inc. Press* .

Dreuw, P., Rybach, D., Deselaers, T., Zahedi, M. & Ney, H. (2007). Speech recognition techniques for a sign language recognition system, *In Proceedings of Interspeech* pp. 2513–2516.

Fleetwood, E. & Metzger, M. (1998). Cued language structure: An analysis of cued american english based on linguistic principles, *Calliope Press, Silver Spring, MD (USA), ISBN 0-9654871-3-X* .

Gibert, G., Bailly, G., Beautemps, D., Elisei, F. & Brun, R. (2005). Analysis and synthesis of the 3d movements of the head, face and hand of a speaker using cued speech, *Journal of Acoustical Society of America* vol. 118(2): 1144–1153.

Gillick, L. & Cox, S. (1989). Some statistical issues in the comparison of speech recognition algorithms, *in Proceedings of ICASSP89* pp. 532–535.

Hennecke, M. E., Stork, D. G. & Prasad, K. V. (1996). Visionary speech: Looking ahead to practical speechreading systems, *in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer* p. 331349.

Heracleous, P., Abboutabit, N. & Beautemps, D. (2009). Lip shape and hand position fusion for automatic vowel recognition in cued speech for french, *in IEEE Signal Processing Letters* 16: 339–342.

Heracleous, P., Kaino, T., Saruwatari, H. & Shikano, K. (2007). Unvoiced speech recognition using tissue-conductive acoustic sensor, *EURASIP Journal on Advances in Signal Processing* 2007.

Heracleous, P., Nakajima, Y., Lee, A., Saruwatari, H. & Shikano, K. (2004). Non-audible murmur (nam) recognition using a stethoscopic nam microphone, *in Proceedings of Interspeech2004-ICSLP* pp. 1469–1472.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G. & Stone, M. (2008). Phone recognition from ultrasound and optical video sequences for a silent speech interface, *in Proc. of Interspeech* pp. 2032–2035.

Jou, S. C., Schultz, T. & Waibel, A. (2004). Adaptation for soft whisper recognition using a throat microphone, *in Proceedings of Interspeech2004-ICSLP* .

Jou, S., Schultz, T., Walliczek, M., Kraft, F. & Waibel, A. (2006). Towards continuous speech recognition using surface electromyography, *in Proc. of ICSLP* pp. 573–576.

Junqua, J.-C. (1993). The lombard reflex and its role on human listeners and automatic speech recognizers, *J. Acoust. Soc. Am.* 1.

Leybaert, J. (2000). Phonology acquired through the eyes and spelling in deaf children, *Journal of Experimental Child Psychology* 75: 291–318.

Montgomery, A. A. & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance, *Journal of the Acoustical Society of America* 73 (6): 2134–2144.

Nakajima, Y., Kashioka, H., Shikano, K. & Campbell, N. (2003). Non-audible murmur recognition, *in Proceedings of EUROSPEECH* pp. 2601–2604.

Nakajima, Y., Kashioka, H., Shikano, K. & Campbell, N. (2005). Remodeling of the sensor for non-audible murmur (nam), *in Proceedings of Interspeech2005-EUROSPEECH* pp. 389–392.

Nakamura, K., Toda, T., Nakajima, Y., Saruwatari, H. & Shikano, K. (2008). Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments, *in Proc. of Intespeech* pp. 2209–2212.

Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C. & Murphy, K. (2002). A coupled hmm for audio-visual speech recognition, *in Proceedings of ICASSP 2002* .

Nicholls, G. & Ling, D. (1982). Cued speech and the reception of spoken language, *Journal of Speech and Hearing Research* 25: 262–269.

O.Cornett, R. (1967). Cued speech, *American Annals of the Deaf* 112: 3–13.

Ong, S. & Ranganath, S. (2005). Automatic sign language analysis: A survey and the future beyond lexical meaning, *IEEE Trans. PAMI* vol. 27, no. 6: 873891.

Potamianos, G., Gravier, G., Garg, A., Cooley, A. S. & Tukey, J. W. (2003). Recent advances in the automatic recognition of audiovisual speech, *in Proc. of the IEEE* 91, Issue 9: 1306–1326.

Revéret, L., Bailly, G. & Badin, P. (2000). Mother: a new generation of talking heads providing a flexible articulatory control for video-realistic speech animation, *in Proceedings of ICSLP* pp. 755–758.

Revéret, L. & Benoît, C. (1998). A new 3d lip model for analysis and synthesis of lip motion in speech production, *in Proceedings of AVSP* .

Toda, T. & Shikano, K. (2005). Nam-to-speech conversion with gaussian mixture models, *in Proc. of Interspeech* pp. 1957–1960.

Tran, V. A., Bailly, G., Loevenbruck, H. & Jutten, C. (2008). Improvement to a nam captured whisper-to-speech system, *in Proc. of Interspeech* pp. 1465–1468.

Uchanski, R. M., Delhorne, L. A., Dix, A. K., Braida, L. D., Reedand, C. M. & Durlach, N. I. (1994). Automatic speech recognition to aid the hearing impaired: Prospects for the automatic generation of cued speech, *Journal of Rehabilitation Research and Development* vol. 31(1): 20–41.

Walliczek, M., Kraft, F., Jou, S.-C., Schultz, T. & Waibel, A. (2006). Sub-word unit based non-audible speech recognition using surface electromyography, *in Proceedings of Interspeech2006-ICSLP* pp. 1487–1490.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2001). The htk book, *Cambridge University Engineering Department* .

Zheng, Y., Liu, Z., Zhang, Z., Sinclair, M., Droppo, J., Deng, L., Acero, A. & Huang, Z. (2003). Air- and bone-conductive integrated microphones for robust speech detection and enhancement, *in Proceedings of ASRU* pp. 249–253.

Speech and Language Technologies

Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

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

IntechOpen