Confusability of Phonemes Grouped According to their Viseme Classes in Noisy Environments

Patrick Lucey, Terrence Martin and Sridha Sridharan

Speech and Audio Research Laboratory
Queensland University of Technology
GPO Box 2434, Brisbane 4001, Australia.
{p.lucey, tl.martin, s.sridharan}@qut.edu.au

Abstract

Using visual information, such as lip shapes and movements, as the secondary source of speech information has been shown to make speech recognition systems more robust to problems associated with environmental noise, training/testing mismatch and channel and speech style variations. Research into utilising visual information for speech recognition has been ongoing for 20 years, however over this period, a study into which visual information is the most useful or pertinent to improving speech recognition systems has yet to be performed. This paper presents a study to determine the confusability of the phonemes grouped into their viseme classes over various levels of noise in the audio domain. The rationale behind this approach is that by establishing the interclass confusion for a group of phonemes in their viseme class, a better understanding can be obtained on the complementary nature of the separate audio and visual information sources and this can be subsequently applied in the fusion stage of an *audio-visual speech processing* (AVSP) system. The experiments performed show high interclass confusion variability at the 0dB and -6dB SNR levels. Further analysis found that this was mainly due to a phonetic imbalance in the dataset. Due to this result, it was suggested that it would be appropriate for an AVSP system used for digit recognition applications heavily weight the visual modality for the phonemes that are most prevalent such as the phoneme *N*.

1. Introduction

Traditionally, speech processing has been thought of as a single sense input in the auditory domain. As a consequence, the majority of the past research on speech processing has been strictly confined to the audio modality. However, the performance of acoustic-based speech processing systems has yet to reach the level required for them to be used for the vast majority of conceivable applications. This is mainly due to the fact that current acoustic-based systems are quite susceptible to environmental noise, training/testing mismatch and channel and speech style variations (Potamianos, Neti, Luettin, and Matthews 2004). Using visual information, such as lip shapes and movements, as the secondary source of speech information has been shown to make speech processing systems more robust to these problems. The field of speech processing which uses both audible and visible speech cues is known as *audio-visual speech processing* (AVSP).

One of the aims of researching AVSP is to emulate the ways in which humans jointly use audio-visual information for speech communication. It is believed that computers can make effective use of the joint audio-visual information for tasks such as speech recognition. As *automatic speech recognition* (ASR) promises to be an integral part of the future human-computer interface (Potamianos, Neti, Gravier, Garg, and Senior 2003), AVSP can increase the accuracy of ASR systems, and aid processes such as out of vocabulary rejection and key word spotting as well as alleviating the problems described previously. Since the visual (video) features are robust to the mismatched condition arising in audio features (i.e audio noise does not affect the video information), fusion of the two sources of information should intuitively result in improvement in performance of speech recognition systems. Many applications such as human-computer interaction; hands-off operation of equipment (photocopiers, vehicles, military aircraft etc.); video conferencing; and speech transcription would greatly benefit from the use of an AVSP system (Chibelushi, Deravi, and Mason 2002).

Research on the topic of AVSP has been ongoing for 20 years, with the first work appearing under the heading of lip reading in 1984 (Petajan 1984). AVSP has been shown to improve ASR systems over this period but the extent to which visual information contributes to this improvement has yet to be explored. The lack of research into this area has motivated the work for this paper.

The basic unit of acoustic speech is called the *phoneme* (Rabiner 1989), which is the theoretical unit for describing the linguistic meaning of speech. Phonemes have the property that if one is replaced by another one, the meaning of the utterance is changed. Similarly, in the visual domain, the basic unit of mouth movements is called a *viseme* (Chen 1998). A generic facial image or lip shape is associated with a viseme in order to represent a particular sound. Strictly speaking, instead of a still image, a viseme can be a sequence of several images that capture the movements of the mouth. However, most visemes can be approximated by stationary images (Chen 1998).

In this paper, a study was conducted to determine the confusability of phonemes grouped into their viseme classes when subjected to various levels of noise in the au-

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

dio domain. The rationale behind this approach was that if it could be determined that a certain group of phonemes in their viseme class were more or less easily recognised then others, then more emphasis or appropriate weighting could be assigned to those phonemes in the audio-visual fusion stage of the AVSP system. Take, for example, the phonemes *M* and *N*. Acoustically these phonemes are easily confused, however, in the visual domain, these two sounds look quite different. This is a good example of the usefulness of the visual modality when trying to recognise acoustically confused sounds. This is because the visual modality provides complementary information which allows a system to recognise these sounds with more accuracy due to the extra information source.

If a phoneme had high or low confusion with other phonemes in the same viseme class (intraclass confusion) in the auditory domain, there would be no real need to use visual information as these sounds look the same in the visual domain and thus does not provide any complementary or useful information about those sounds. However, if a phoneme were to be confused with another phoneme which was in a different viseme class (interclass confusion), then visual information would be complementary to the information in the audio domain. This is because these sounds look different in the visual domain. As obtaining the visual information from a speaker's lip is computationally expensive and requires vast amounts of memory, knowing which sounds require visual information has the potential to greatly save the amount of memory necessary for an AVSP system as well as improving the speed of the system. The main aim of this study was to determine the intraclass and interclass confusion of the phonemes in the audio domain under different noisy environments so as to give a better understanding on what type of visual information would be useful in an AVSP system. As a result of this study, it is hoped that weights can be assigned to specific phonemes and visemes according to the visual and audio stream reliabilities in the fusion stage. This type of fusion is known as *adaptive fusion*. It is hoped that the novel adaptive fusion method mentioned in this paper will help improve speech recognition results. This study is in contrast to most of the work performed on adaptive fusion in AVSP, which has basically concerned itself with assigning weights to either the entire acoustic or visual domain depending on how reliable the audio signal is by estimating the the audio signal-to-noise ratio (SNR) (Heckmann, Berthommier, and Kroschel 2002; Glotin, Vergyri, Neti, Potamianos, and Luettin 2001).

The rest of this paper is organised as follows. In Section 2, a description of how the phonemes are mapped into their viseme counterparts is given. Section 3 details the CUAVE audio-visual database that was used in the experiments. This section also documents how the confusability of the phonemes based on their visemes classes were found. Section 4 gives the results from the experiments and also discusses their relevance. Finally, Section 5 closes with some observations on the intraclass and interclass confusions that were found from this study and also suggests the possible areas of research that this particular study might lead to.

2. Phoneme to Viseme Mapping

For English, the ARPABET table, consisting of 48 phonemes, is commonly used to classify phonemes (Rabiner and Juang 1993). However, currently there is no standard viseme table used by all researchers. As a result, a viseme table for this study was based on the work performed by Lee and Sook (2002). This was used in collating a table which effectively mapped all the possible phonemes to visemes. This is given in Table 1. This table shows the typical 48 phonemes used in English language including silence and short pauses, grouped into their 14 viseme classes.

Table 1: Phoneme to viseme mapping

Phoneme	Viseme	Phoneme	Viseme
P	/p/	K	/k/
B		G	
M		N	
EM		L	
F	/f/	NX	
V		HH	
T	/t/	Y	
D		EL	
S		EN	
Z		IY	/iy/
TH		IH	
DH		AA	/aa/
DX		AH	/ah/
W	/w/	AX	
WH		AY	
R		ER	/er/
CH	/ch/	AO	
JH		OY	
SH		IX	
ZH		OW	
EH	/ey/	UH	/uh/
EY		UW	
AE		SIL	/sp/
AW		SP	

As can be seen from Table 1, many acoustic sounds are visually ambiguous, and accordingly different phonemes can be classed using the same viseme. There is therefore a many-to-one mapping between phonemes and visemes. By the same token there are many visemes that are acoustically ambiguous. An example of this can be seen in the acoustic domain when people spell words on the telephone, expressions such as 'B as in boy' or 'D as in David' are often used to clarify such acoustic confusion. These confusion sets in the auditory modality are usually distinguishable in the visual modality (Chen 2001). This highlights the bimodal nature of speech and the fact that to properly understand what is being said, information is required from both modalities. The extra information contained in the visual modality can be used to improve standard speech processing applications such as speech recognition. The bimodal nature of speech is also illustrated by the McGurk effect (McGurk and MacDonald 1976). The McGurk effect shows that when humans are presented with conflict-

Accepted after full paper review

ing acoustic and visual stimuli, the perceived sound may not exist in either modality. For example, when a person hears the sound /ba/, but watches the sound /ga/, the person may not perceive either /ba/ or /ga/ but may perceive /da/.The McGurk effect highlights the requirement for both acoustic and visual cues in the perception of speech.

3. Experimental Setup

3.1. Training and Test Datasets

Training and evaluation speech was taken from the Clemson University, *CUAVE*, audio-visual database (Patterson, Gurbuz, Tufekci, and Gowdy 2002). The CUAVE database was selected as it is presently the only common audio-visual database which is available for all universities to use. This is important for benchmarking and comparison purposes. Even though the XM2VTS database (Messer, Matas, Kittler, Luettin, and Maitre 1999) is also another database which is available to researchers for the same purposes, the CUAVE database was chosen due to the fact it is freely available. The CUAVE database consists of two major sections, one of individual speakers and one of speakers pairs. For this study, only the individual speakers were used. The part with individual speakers consists of 36 speakers, 19 male and 17 female.

Even though the audio modality was only used in this study, it was deemed important to use a audio-visual database as the intention is to use these results in a future study using an AVSP system. Due to the difficulties associated with the high volumes of data necessary for simultaneous video audio, the creation of audio-visual databases has been limited. As a result, the CUAVE database is a speaker-independent corpus of over 7 000 utterances of only connected and isolated digits (0-9).

As connected and isolated digits were the only words spoken, the phoneme to viseme mapping was a subset of the entire set given in Table 1. From the original 48 phonemes and 14 visemes, only 22 phonemes and 10 visemes were required. The simplified mapping is shown in Table 2.

Each speaker in the CUAVE database was recorded speaking digits in several different styles. Initially, 100 isolated digits were spoken. 60 connected digits including telephone-number-like sequences were then spoken. The database was recorded in an isolated sound booth, using a MiniDV camera. Several microphones were tested. An on camera microphone produced the best results: audio that was clear from clicking or popping due to speaker movement and video where the microphone did not block the view of the speaker (Patterson, Gurbuz, Tufekci, and Gowdy 2002). From the recordings, only disruptive mistakes were removed, but occasional vocalised pauses and mistakes in speech were kept for realistic test purposes.

The training and test data sets used for this study were based on the CUAVE database specification. 30 subjects totalling 1.25 hours of audio data were used for training and 6 subjects totalling 0.25 hours were used for testing. Additive Gaussian noise was added to the test data at various SNR levels. These levels were:

Table 2: Phoneme to viseme mapping for digit recognition

Phoneme	Viseme
F	/f/
V	
T	/t/
S	
Z	
TH	
W	/w/
R	
K	/k/
N	
IY	/iy/
IH	
EH	/eh/
EY	
AH	/ah/
AY	
AO	/ao/
OW	
UW	/uh/
SIL	/sp/
SP	

- clean speech
- 18dB
- 12dB
- 6dB
- 0dB
- -6dB

The phoneme/viseme recogniser was trained on clean speech and tested under noisy conditions.

3.2. Phoneme Recognition based on Viseme Classes

For this study, determining the confusability of the phonemes grouped by their viseme classes was essentially a viseme recognition problem. This is due to the fact that the intraclass and interclass confusions based on the viseme classes were the results that were required for analysis. The viseme recogniser used was based on a phoneme recogniser. In this approach, the audio signals were coded into a sequence of phonemes. The phoneme sequence was mapped to a viseme sequence using Table 2. The diagram of this approach is illustrated in Figure 1. In Figure 1, it can be seen that the phonemes are modelled as Hidden Markov Models (HMMs).

An HMM is a stochastic model, into which some temporal restrictions can be incorporated. It can be used to capture the acoustic characteristic of a speech sound. An HMM can be considered as a special case of the Bayesian classifier, where the most probable token sequence $\hat{U}$ for given speech X is selected among all possible token sequences U^* as follows;

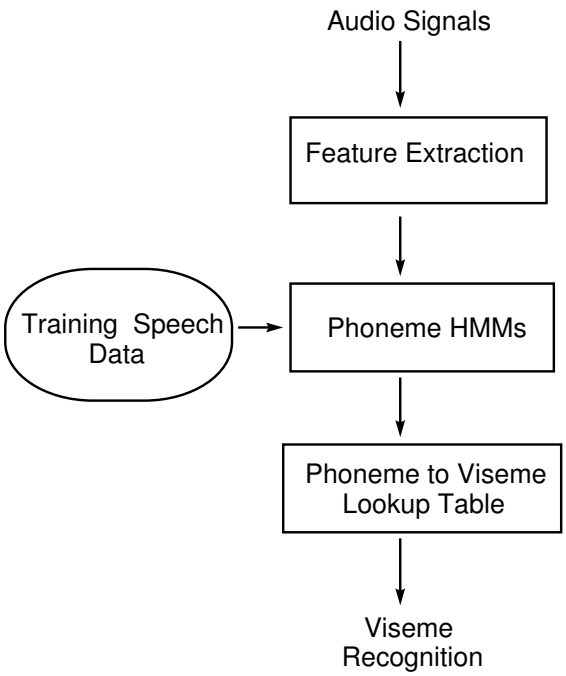

Figure 1: A viseme recognition system using phoneme HMMs

$$\hat{U} = \arg \max_{U \in U^*} P(U|X) \quad (1)$$

One of the distinguishing characteristics of speech is that it is dynamic. Even within a small segment such as a phoneme, the speech sound changes gradually. The previous phones affect the beginning of a phone, the middle portion of the phone is relatively stable, and the following phones affect the end of the phone. The temporal information of speech feature vectors plays an important role in the recognition process. In order to capture the dynamic characteristics of speech within the framework of the Bayesian classifier, certain temporal restrictions should be imposed. A 5 state left to right HMM is sometimes used for this purpose (S. Lee and D. Yook 2002). This model architecture was used for this study.

4. Phoneme Confusability Results

The data used in these experiments were parameterised using Perceptual Linear Prediction (PLP) coefficient feature extraction. Each feature vector used for the experiments was composed of 12 dimensional PLPs, normalised energy, and their first and second order time derivatives, resulting in a 39 dimensional vector. It was computed every 10 milliseconds using 25 milliseconds long Hamming windowed speech signals. The HMMs were trained using 7 000 utterances from the CUAVE training data. For the phoneme recogniser, 22 HMMs including silence models are created during training. All HMMs were modelled using 5 state left-to-right HMM with 8 Mixtures per state. As the training data from the CUAVE database was relatively small, these models were re-estimated using phone models from the Wall Street Journal (WSJ) database (Baker and Paul 1992). This was done as the WSJ database had much more data than the CUAVE database (nearly 80 hours compared to 1 hour). This resulted in far superior results than the flat-start phone models that were obtained using the CUAVE database.

Table 3: Intraclass recognition rates (%) of phonemes grouped according to their viseme classes over various levels of noise

Viseme	clean	18dB	12dB	6dB	0dB	-6dB
/ah/	95.6	97.8	97.8	97.6	88.7	32.4
/ao/	96.2	96.7	95.9	80.5	55.3	37.0
/eh/	97.4	97.3	96.2	87.6	62.9	15.1
/f/	97.5	97.2	97.8	92.6	60.9	33.3
/iy/	98.7	98.5	97.5	92.7	60.7	18.0
/k/	97.0	96.3	96.9	95.8	87.3	48.0
/t/	99.1	98.9	95.0	86.0	56.9	35.0
/uh/	97.8	95.6	97.8	94.4	35.4	17.7
/w/	93.7	94.4	92.9	81.7	69.9	33.6
/sp/	93.9	94.4	94.3	95.8	96.8	100.0

The results from the viseme recogniser over the various levels of noise are shown in Table 3. From this table it can be seen that the recognition rates for all the visemes were very high, especially at the clean speech, 18dB and 12dB levels, which suggests low interclass confusion. The results at a SNR of 6dB also gave relatively good recognition rates. However, at the 0dB and -6dB SNR levels, the recognition rates had high variability across the viseme classes. This suggests that there was high interclass confusion most likely due to the excessive noise levels present in the audio signal.

Looking at the intraclass confusion matrix of the phonemes grouped according to their visemes at 0dB in Table 4, it can be seen that there is variability in the interclass confusion of the viseme classes. For the viseme classes /ah/ and /k/, the phonemes in these classes exhibit low interclass confusion. It would appear on the surface, that the phonemes in these classes would require little visual information as no extra information would be introduced from the visual modality as these phonemes look the same. However, upon further investigation, the phonemes in these groups did not share the same commonalities, i.e. the intraclass recognition rates were greatly varied. For example, the viseme class /k/, is made up of the phonemes *N* and *K*. The phoneme recognition rate of *N* was 85.2% which was similar to the recognition rate for its corresponding viseme class. However the phone *K* only had a recognition rate of 9.4%, which is vastly different. Yet when these rates were

merged into their viseme class, a very high recognition rate was achieved. This was because there was approximately 15 times the examples for the phoneme *N* uttered in the test set than *K*, which introduced massive bias into the results. This was observed to be a common trend throughout the result obtained.

Table 4: Intraclass confusion matrix at 0dB, showing the percentage of the phonemes in their viseme classes vertically being identified as phonemes within the viseme classes on the horizontal

	/ah/	/ao/	/eh/	/f/	/iy/	/k/	/t/	/uh/	/w/	/sp/
/ah/	88.7	3.2	-	-	1.6	-	1.6	1.6	-	3.2
/ao/	10.1	55.3	-	1.3	-	27.7	-	5.0	0.6	-
/eh/	-	11.3	62.9	-	8.2	10.3	-	3.1	4.1	-
/f/	2.4	6.8	0.4	60.6	-	12.4	-	2.0	14.1	1.2
/iy/	2.0	6.1	0.3	0.3	60.7	22.0	0.3	4.4	1.0	2.7
/k/	1.4	4.3	0.5	3.3	-	87.3	0.3	1.1	1.9	-
/t/	0.7	3.6	-	6.2	1.1	23.4	56.9	2.2	2.2	3.6
/uh/	-	2.4	-	-	-	61.0	1.2	35.4	-	-
/w/	4.1	1.0	-	2.0	0.5	7.7	-	14.3	69.9	0.5
/sp/	0.6	0.2	-	0.2	-	1.4	0.1	0.5	0.1	96.8

This problem highlights a major problem in AVSP. As audio-visual databases require vast amounts of data to accommodate both audio and video modalities, these databases have been only available for small vocabulary tasks such as digit recognition. The problem of databases which are designed for isolated and connected digit recognition, are that they are not phonetically balanced like the TIMIT database (NIST Speech Disc 1-1.1 1990), i.e. it does not have equal representation of all phonemes spoken therefore introducing bias as seen in this study. At the present stage, there is currently only one phonetically balanced audio-visual database and it is designed for large vocabulary continuous speech recognition. This database is produced by IBM (Neti, Potamianos, Luettin, Matthews, Glotin, and Vergyri 2001), but is not commercially available.

Also at the 0dB SNR ratio, the viseme class /uh/ had high confusability with a recognition rate of 35.4%. But looking at Table 4, it can be seen that there was high interclass confusability with viseme /k/, with the phonemes in /uh/ being confused with the phonemes in /k/ over 60% of the time. This can also be attributed to the dataset problem as there is 4 times as much data for the phonemes in the viseme class /k/ compared to the phonemes which were in the viseme class /uh/. As can be seen in Table 4, this type of interclass confusion was present for the majority of the other viseme classes.

Table 5: Intraclass confusion matrix at -6dB, showing the percentage of the phonemes in their viseme classes vertically being identified as phonemes within the viseme classes on the horizontal

	/ah/	/ao/	/eh/	/f/	/iy/	/k/	/t/	/uh/	/w/	/sp/
/ah/	32.4	-	-	-	-	-	-	-	-	67.6
/ao/	2.5	37.0	-	-	-	58.8	0.8	0.8	-	-
/eh/	1.9	5.7	15.1	-	-	35.8	-	-	1.9	39.6
/f/	1.3	1.3	-	33.3	-	38.7	-	-	0.7	24.7
/iy/	2.2	1.3	1.8	0.9	18.0	25.0	-	0.4	-	50.4
/k/	-	0.5	-	0.5	-	98.0	-	0.2	0.2	0.5
/t/	2.6	3.0	-	3.4	0.4	37.2	35.0	1.7	1.3	15.4
/uh/	3.2	6.5	1.6	4.8	3.2	59.7	-	17.7	3.2	-
/w/	-	0.9	-	0.9	-	16.8	-	1.9	33.6	45.8
/sp/	-	-	-	-	-	-	-	-	-	100

Another interesting result stemming from the experiments performed in this work showed that when the noise level is extreme (i.e. $\leq$ -6dB), a significant proportion of phonemes are confused with silences and short pauses. This can be seen in Table 5. Upon reflection, this result is quite intuitive, as in very noisy environments it is quite difficult to decipher what sound is being made, so people look to visual modality for complementary information. This is backed up by Heckmann et al's (2002) work, as in their study they found at the -6dB SNR level, 75% of the phonemes got confused with silences and short pauses. However, in the visual domain only 22% of the visemes had been confused with these pauses.

5. Conclusions

AVSP is becoming a very important area of research as it has the potential to make speech recognition systems tractable for real-world applications. In this paper, the results for a study conducted to determine the confusability of the phonemes grouped into their visemes classes over various levels of noise were shown. The results showed that there was low interclass confusion of the phonemes in their viseme classes at the clean speech, 18dB and 12dB levels SNR levels. The results at 6dB were also quite good. At the 0dB and -6dB level, the results displayed that there was high variability in the amount of interclass confusion across some of the viseme classes. Upon further investigation it was shown that the confusion was due to the dataset being not phonetically balanced with some phonemes being uttered almost 15 times more than other phonemes.

Due to this dataset inequality, it would be advisable for an AVSP system being implemented for a digit recognition application to pay particular attention to the phonemes

which were being uttered the most. This is due to the fact that these particular phonemes are the sounds being uttered the most by a considerable factor. If obtaining and using the visual information of a speaker's lip proves to be too computationally expensive to implement in a digit recognition application, just using the visual information on these phonemes, may be a worthwhile exercise. For example, in this study, it was found that the phoneme *N* was uttered approximately 15 times more than the other phonemes. Just having an AVSP system focus on this particular phoneme may improve the recognition rate when comparing it to a ASR system.

Also at these very noisy audio levels, it would also be wise to have an AVSP system able to segment speech effectively due to the amount of phonemes being confused with silences and short pauses. As visual speech not only gives information about speech itself, it also gives segmentation information. In future work, it is hoped to use the visual information to select non-speech segments for SNR estimation in the audio channel to help in assigning weights to the various phonemes in an AVSP system. It is also planned to continue this study, but this time, we intend to study the confusability of visemes in the visual domain. Also it is intended that we study the confusability of phonemes according to their viseme classes on a large vocabulary continuous speech recognition audio-visual database.

6. Acknowledgements

We would like to thank Clemson University for freely supplying us their CUAVE audio-visual database for our research.

References

Baker, J. M. and D. B. Paul (1992). The design for the wall street journal-based csr corpus. *ICSLP*.

Chen, T. (1998, May). Audio-visual integration in multimodal communication. *Proceedings of the IEEE 86*, 837–852.

Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, 9–31.

Chibelushi, C., F. Deravi, and J. Mason (2002). A review of speech-based bimodal recognition". *IEEE Trans. Multimedia 4*(1), 23–37.

Glotin, H., D. Vergyri, C. Neti, G. Potamianos, and J. Luettin (2001). Weighting schemes for audio-visual fusion in speech recognition. In *ICASSP*, Salt Lake City, Utah.

Heckmann, M., F. Berthommier, and K. Kroschel (2002). Noise adaptive stream weighting in audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing 2002*(11), 1260–1273.

McGurk, H. and J. MacDonald (1976, December). Hearing lips and seeing voices. *Nature*, 746–748.

Messer, K., J. Matas, J. Kittler, J. Luettin, and G. Maitre (1999). Xm2vts: The extended m2vts database. In *Proceedings of the International Conference on Audio and Video-based Biometric Person Authentication*, Washington D.C., pp. 72–76. IEEE.

Neti, C., G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri (2001, October 3-5). Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop. In *Workshop on Multimedia Signal Processing, Special Section on Joint Audio-Visual Processing*, Cannes.

NIST Speech Disc 1-1.1 (1990). Timit acoustic-phonetic continuous speech corpus.

Patterson, E. K., S. Gurbuz, Z. Tufekci, and J. N. Gowdy (2002). Cuave: a new audio-visual database for multimodal human-computer interface research. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando.

Petajan, E. (1984). Automatic lipreading to enhance speech recognition. In *IEEE Global Telecommunications Conference*, Atlanta, GA, USA, pp. 265–272. IEEE.

Potamianos, G., C. Neti, G. Gravier, A. Garg, and A. W. Senior (2003). Recent advances in the automatic recognition of audio-visual speech. *Proc. of the IEEE 91*(9).

Potamianos, G., C. Neti, J. Luettin, and I. Matthews (2004). Audio-visual automatic speech recognition: An overview. In G. Bailly, E. Vatikiotis-Bateson, and P. Perrier (Eds.), *Issues in Visual and Audio-Visual Speech Processing*. Boston: MIT Press.

Rabiner, L. and B. Juang (1993). *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 257–286.

S. Lee and D. Yook (2002). Audio-to-visual conversion using hidden markov models. In *Proceedings of the 7th Pacific Rim International Conference on Artificial Intelligence*, pp. 563–570. Springer-Verlag.

Accepted after full paper review