AUTOMATIC SPEECHREADING OF IMPAIRED SPEECH

Gerasimos Potamianos and *Chalapathy Neti*

IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
E-mails: {gpotam,cneti}@us.ibm.com

ABSTRACT

We investigate the use of visual, mouth-region information in improving automatic speech recognition (ASR) of the speech impaired. Given the video of an utterance by such a subject, we first extract appearance-based visual features from the mouth region-of-interest, and we use a feature fusion method to combine them with the subject's audio features into bimodal observations. Subsequently, we adapt the parameters of a speaker-independent, audio-visual hidden Markov model, trained on a large database of hearing subjects, to the audio-visual features extracted from the speech impaired videos. We consider a number of speaker adaptation techniques, and we study their performance in the case of a single speech impaired subject uttering continuous read speech, as well as connected digits. For both tasks, maximum-a-posteriori adaptation followed by maximum likelihood linear regression performs the best, achieving a word error rate relative reduction of 61% and 96%, respectively, over unadapted audio-visual ASR, and a 13% and 58% relative reduction over audio-only speaker-adapted ASR. In addition, we compare audio-only and audio-visual speaker-adapted ASR of the single speech impaired subject to ASR of subjects with normal speech, over a wide range of audio channel signal-to-noise ratios. Interestingly, for the small-vocabulary connected digits task, audio-visual ASR performance is almost identical across the two populations.

1. INTRODUCTION

Visual information in the speaker's mouth region is known to benefit both human *speech perception* [1] and *automatic speech recognition* (ASR) [2]. Indeed, the visual modality role in speech intelligibility in noise has been quantified as early as in 1954 [3], whereas the fusion of audio and visual stimuli by humans has been shown by the *McGurk effect* [4]. More recently, incorporating visual information in ASR systems (also known as *automatic speechreading*) has resulted in improved speech recognition in a variety of audio channel conditions and for a number of recognition tasks, initially limited to isolated or connected words [5]–[7], and later demonstrated in the *speaker-independent* (SI),

large-vocabulary, continuous speech recognition (LVCSR) domain [8], [9].

The use of visual speech information is of particular importance to the *hearing impaired*. Deaf people can speechread well, and possibly better than the general population [10]. Furthermore, mouth movement plays an important role in both sign language and simultaneous communication between the deaf [11]. It is plausible that such visual speech information can be recognized by an automatic system, thus helping transcribe speech produced by the hearing impaired, and improving accessibility and communication for the deaf.

In this paper, we are interested in automatically recognizing mouth movements as they are generated during *impaired speech* in simultaneous communication by subjects with profound hearing loss during the period of normal language acquisition, and in augmenting automatic recognition of impaired speech by such visual information. This of course corresponds to the traditional automatic speechreading task [5], with the main difference being the severely degraded nature of the audio channel speech information. One however hopes that the visual speech information is significantly less degraded, benefiting impaired ASR possibly more than in the normal speech case.

To study the use of visual speech information in ASR for the speech impaired, we collect audio-visual speech by a single such subject, uttering continuous read speech (large-vocabulary) and connected digit strings (small-vocabulary task). We then employ the automatic speechreading system reported in [9] to extract audio-visual features from the subject video sequences, and we use a *hidden Markov model* (HMM) classifier [12] to automatically recognize speech based on such features. Due to the small amount of impaired speech data (insufficient to reliably train HMMs), we use *speaker adaptation* of HMMs, that have been trained on an appropriate speaker-independent, large-vocabulary audio-visual database of subjects with normal speech, to the impaired speech audio-visual data. Speaker adaptation is traditionally used in practical audio-only ASR systems to improve speaker-independent system performance, when little data from a speaker of interest are available [13]–[15],
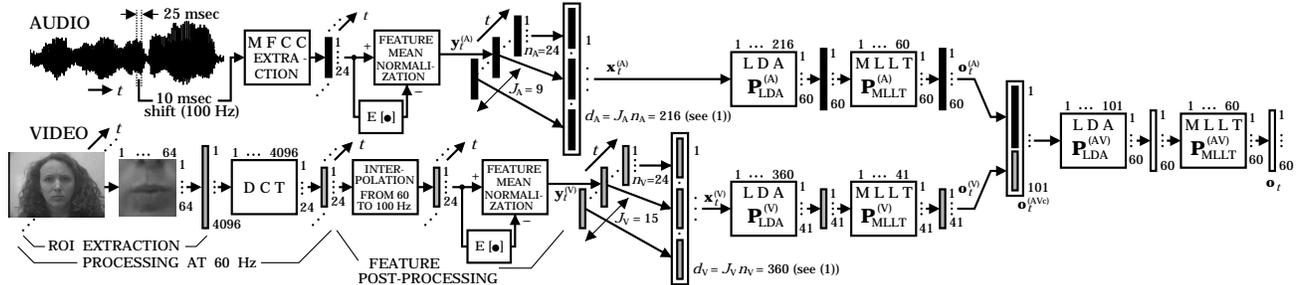
Figure 1: Feature extraction for audio-visual ASR by a hierarchical, two-stage application of LDA and MLLT (from [9]).

and has been successfully applied to audio-visual ASR in [16]. Various algorithms exist for HMM parameter adaptation, such as *maximum-a-posteriori* (MAP) adaptation [13], [14], and *maximum likelihood linear regression* (MLLR) [14], [15]. In this work, we consider both methods, and we demonstrate that their combination outperforms either one. Furthermore, we investigate feature level (front end) adaptation. We report that audio-visual adaptation significantly improves automatic impaired speech recognition over both audio-only adapted, as well as audio-visual unadapted ASR, for both large- and small-vocabulary tasks considered. In addition, a comparison of audio-visual ASR between impaired and normal speech reveals almost identical performance on the connected digits task.

The paper is structured as follows: Section 2 reviews the automatic speechreading system used to extract audio-visual features and to recognize speech. Section 3 is devoted to an overview of the speaker adaptation techniques, employed in this paper to improve audio-visual ASR of the speech impaired. Section 4 describes the audio-visual databases, and Section 5 reports our experimental results. Finally, Section 6 summarizes the paper.

2. AUTOMATIC SPEECHREADING

Various automatic speechreading systems have appeared in the literature over the last two decades. Three main factors differentiate such systems [5]: (a) The choice of visual features (mouth region *appearance* versus *shape* based features); (b) The integration of audio and visual features into a bimodal speech classifier (*feature* versus *decision* fusion); and (c) The speech classifier considered (an HMM versus a neural network based system, for example). In this work, we use the automatic speechreading system reported in [9], that employs mouth region appearance based visual features, hierarchical discriminant feature fusion, and HMMs for speech classification (see also [8]). The system is briefly described below.

2.1. Audio- and Visual-Only Features

Given an utterance, video, audio- and visual-only *static* features are first extracted. The former ones, denoted by $\mathbf{y}_t^{(A)} \in \mathbb{R}^{n_A}$, consist of 24 mel-frequency cepstral coefficients, computed over a sliding window of 25 msec, at a rate of 100 Hz, followed by the application of *feature mean normalization* (FMN) [12]. To extract static visual features, a statistical face tracking algorithm is first used to detect the speaker's face and estimate the mouth location and size [17]. Based on these, a size-normalized, 64×64 pixel *region-of-interest* (ROI) is obtained for every video frame at 60 Hz, containing the speaker's mouth. Subsequently, a two-dimensional, separable, *discrete cosine transform* (DCT) is applied to the ROI, and the 24 highest-energy DCT coefficients are retained as features. To facilitate audio-visual fusion, *linear interpolation* is employed to obtain visual features, time-synchronous to the audio ones at 100 Hz. Finally, FMN is used to compensate for lighting variations, providing the final visual-only static features $\mathbf{y}_t^{(V)} \in \mathbb{R}^{n_V}$ (see Figure 1).

To obtain *dynamic* audio- and visual-only features, J_s ($s = $ A ,V) consecutive static feature vectors are concatenated into vectors

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

of dimension $d_s = J_s\, n_s$. Subsequently, vectors (1) are projected onto a lower D_s-dimensional space by means of a *linear discriminant analysis* (LDA) [9] based projection matrix $\mathbf{P}_{\mathrm{LDA}}^{(s)}$, that improves discrimination among a set of classes of interest, $\mathcal{C}$ (here, the HMM states; see also (5)). The resulting vectors are further "rotated" by means of a *maximum likelihood linear transformation* (MLLT) [9] matrix $\mathbf{P}_{\mathrm{MLLT}}^{(s)}$, that improves data maximum likelihood modeling, under the assumption of data class-conditional Gaussian probability densities with diagonal covariances. The final audio- and visual-only features of dimension D_s are

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

Values $n_A = 24$, $J_A = 9$, $D_A = 60$, and $n_V = 24$, $J_V = 15$, $D_V = 41$, are used [9].

2.2. Audio-Visual Feature Fusion

The joint, concatenated audio-visual feature vector is

$$\mathbf{o}_t^{(\mathrm{AVc})} = [\, \mathbf{o}_t^{(\mathrm{A})\,\top},\, \mathbf{o}_t^{(\mathrm{V})\,\top}\,]^{\top} \in \mathrm{R}^{D_{\mathrm{AV}}}, \qquad (3)$$

where $D_{\mathrm{AV}} = D_{\mathrm{A}} + D_{\mathrm{V}} = 101$. To achieve dimensionality reduction, vectors (3) are projected onto a lower-dimensional space by means of a second stage of LDA and MLLT, giving rise to the final audio-visual features

$$\mathbf{o}_t = \mathbf{P}_{\mathrm{MLLT}}^{(\mathrm{AV})}\, \mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}\, \mathbf{o}_t^{(\mathrm{AVc})} \in \mathrm{R}^{D}, \qquad (4)$$

of dimension $D = 60$.

2.3. Audio-Visual Speech Modeling

The generation of a sequence of features (4) is modeled by a *single-stream* HMM, with *emission* (class conditional observation) probabilities [12], given by

$$Pr[\,\mathbf{o}_t|\,\mathsf{c}\,] = \sum_{k=1}^{K_{\mathsf{c}}} w_{\mathsf{c}k}\, \mathcal{N}_D(\,\mathbf{o}_t;\mathbf{m}_{\mathsf{c}k},\mathbf{s}_{\mathsf{c}k}\,), \qquad (5)$$

and *transition* probabilities $\mathbf{a}_{\mathrm{tr}} = \{Pr[\mathsf{c}'|\,\mathsf{c}''],\ \mathsf{c}',\mathsf{c}'' \in \mathcal{C}\}$. The HMM parameter vector is therefore

$$\mathbf{a} = [\,\mathbf{a}_{\mathrm{tr}},\, (w_{\mathsf{c}k},\mathbf{m}_{\mathsf{c}k},\mathbf{s}_{\mathsf{c}k}),\ k = 1,...,K_{\mathsf{c}},\ \mathsf{c} \in \mathcal{C}\,], \quad (6)$$

where $\mathsf{c} \in \mathcal{C}$ denote the HMM context dependent states (classes), mixture weights $w_{\mathsf{c}k}$ are positive adding to one, K_{c} denotes the number of mixtures, and $\mathcal{N}_D(\mathbf{o};\mathbf{m},\mathbf{s})$ is the D-variate normal distribution with mean $\mathbf{m}$ and a diagonal covariance matrix, its diagonal being denoted by $\mathbf{s}$.

To obtain estimates of (6), we are given, let's say, I audio-visual observation training sequences $\mathbf{O}^{(i)} = [\,\mathbf{o}_{1,i},...,\mathbf{o}_{T_i,i}]$ of duration T_i, $i = 1,...,I$, with the entire training set observations being denoted by $\mathbf{O} = [\mathbf{O}^{(1)},..,\mathbf{O}^{(I)}]$. Let us also denote by $\mathbf{c}(i)$ any HMM state sequence for utterance i. Given two HMM parameter vectors $\mathbf{a}'$, $\mathbf{a}''$, the *auxiliary function* for the *expectation-maximization* (EM) algorithm is [12]

$$Q(\mathbf{a}',\mathbf{a}''|\mathbf{O}) = \sum_{i=1}^{I}\sum_{\mathbf{c}(i)} Pr[\mathbf{O}^{(i)},\mathbf{c}(i)|\mathbf{a}'\,]\log Pr[\mathbf{O}^{(i)},\mathbf{c}(i)|\mathbf{a}''\,].$$

Then, given a current HMM parameter vector at iteration j, $\mathbf{a}^{(j)}$, we obtain a re-estimated parameter vector [12]

$$\mathbf{a}^{(j+1)} = \arg\max_{\mathbf{a}} Q(\mathbf{a}^{(j)},\mathbf{a}\,|\,\mathbf{O}). \qquad (7)$$

In this paper, we use 3 iterations of (7) when training speaker-independent HMMs, or in MAP adaptation (see Section 3.2).

3. AUDIO-VISUAL SPEAKER ADAPTATION

Given few bimodal adaptation data $\mathbf{O}^{(\mathrm{AD})}$ from a particular speaker, and a baseline speaker-independent (SI) HMM (5) with parameters $\mathbf{a}^{(\mathrm{SI})}$ (see (6)), we wish to estimate adapted HMM parameters $\mathbf{a}^{(\mathrm{AD})}$ that better model the audio-visual

observations of the particular speaker. Two popular algorithms for speaker adaptation are maximum likelihood linear regression (MLLR) [14], [15] and maximum-a-posteriori adaptation (MAP) [13], [14]. MLLR obtains a maximum likelihood estimate of a *linear transformation* of the HMM means, while leaving covariance matrices, mixture weights, and transition probabilities unchanged, and it provides successful adaptation with a small amount of adaptation data $\mathbf{O}^{(\mathrm{AD})}$ (rapid adaptation). On the other hand, MAP follows the *Bayesian* paradigm for estimating the HMM parameters, given $\mathbf{O}^{(\mathrm{AD})}$. MAP estimates of HMM parameters slowly converge to their EM-obtained estimates as the amount of training (here, adaptation) data becomes large, however such a convergence is slow, and, therefore, MAP is not suitable for rapid adaptation. In practice, MAP is often used in conjunction with MLLR [14].

3.1. MLLR Adaptation

Let $\mathcal{P}$ be a *partition* (obtained by K-means clustering [12], for example) of the set of all Gaussian mixture components of HMM (5), and let $p \in \mathcal{P}$ denote any member of this partition. Then, we seek MLLR adapted HMM parameters

$$\mathbf{a}^{(\mathrm{MLLR})} = [\,\mathbf{a}_{\mathrm{tr}},\, (w_{\mathsf{c}k},\mathbf{m}_{\mathsf{c}k}^{(\mathrm{MLLR})},\mathbf{s}_{\mathsf{c}k}),$$
$$k = 1,...,K_{\mathsf{c}},\ \mathsf{c} \in \mathcal{C}\,], \quad (8)$$

where the HMM means are linearly transformed as

$$\mathbf{m}_{\mathsf{c}k}^{(\mathrm{MLLR})} = \mathbf{W}_p\, [\,1,\mathbf{m}_{\mathsf{c}k}^{\top}]^{\top}, \qquad (9)$$

where $(\mathsf{c},k) \in p$, and $\mathbf{W}_p$, $p = 1,...,|\mathcal{P}|$ are matrices of dimension $D \times (D+1)$. The transformation matrices are estimated on basis of the adaptation data $\mathbf{O}^{(\mathrm{AD})}$ [15], by means of the EM algorithm solving, similarly to (7),

$$\mathbf{a}^{(\mathrm{MLLR})} = \arg\max_{\mathbf{a}\ \mathrm{satisfy}\ (8),\,(9)} Q(\mathbf{a}^{(\mathrm{SI})},\mathbf{a}\,|\,\mathbf{O}^{(\mathrm{AD})}). \qquad (10)$$

Closed form solutions for the unknown matrices exist, if the HMM covariances are diagonal [15].

3.2. MAP Adaptation

Our MAP implementation is similar to the *approximate* MAP adaptation algorithm (AMAP) [14]. AMAP interpolates the "counts" of the speaker-independent training data and the adaptation data. If $\mathbf{O}^{(\mathrm{SI})}$ denotes the training data observations for the SI HMM, and $\mathbf{O}^{(\mathrm{AD})}$ denotes the adaptation data, we obtain the training data $\mathbf{O}$ of the adapted HMM as

$$\mathbf{O} = [\,\mathbf{O}^{(\mathrm{SI})},\underbrace{\mathbf{O}^{(\mathrm{AD})},...,\mathbf{O}^{(\mathrm{AD})}}_{n\,\mathrm{times}}]. \qquad (11)$$

Then, (7) is used to estimate the adapted HMM parameters, $\mathbf{a}^{(\mathrm{MAP})}$. HMM parameters of all mixtures are adapted, provided the adaptation data contain instances of the mixture component in question. Here, we use $n = 15$, in (11).

Speech condition	Recognition task	Training set			Adaptation set			Test set		
		Utter.	Dur.	Sub.	Utter.	Dur.	Sub.	Utter.	Dur.	Sub.
Normal	LVCSR	17111	34:55	239	855	2:03	26	1038	2:29	26
	DIGITS	5490	8:01	50	670	0:58	50	529	0:46	50
Impaired	LVCSR	N / A			50	0:11	1	50	0:11	1
	DIGITS	N / A			80	0:08	1	60	0:06	1

Table 1: The audio-visual databases considered in this paper and their partitioning into training, adaptation, and test sets (number of utterances, duration (in hours), and number of subjects are depicted for each set). Two recognition tasks are considered: Continuous read speech (LVCSR), and connected digits (DIGITS). HMMs trained on the normal speech LVCSR training set are adapted on the single subject, speech impaired LVCSR and/or DIGITS adaptation sets (see also Tables 2 and 3). Test set performances between the two speech conditions are also compared (see Figures 2 and 3). For fairness, normal speech results are reported after MLLR adaptation, per subject.

3.3. LDA and MLLT Matrix Adaptation

In addition to adapting HMM parameters to a particular subject, one may seek to adapt the front end to better capture the speech information of this subject. For the audio-visual front end presented in Section 2, a simple form of *front end adaptation* is to re-estimate the LDA and MLLT matrices of the single-modality features, $\mathbf{P}_{\mathrm{LDA}}^{(s)}$, $\mathbf{P}_{\mathrm{MLLT}}^{(s)}$, for $s = \mathrm{A},\mathrm{V}$ (see (2)), and/or the audio-visual feature LDA and MLLT matrices $\mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}$, $\mathbf{P}_{\mathrm{MLLT}}^{(\mathrm{AV})}$ (see (4)). Here, we simply compute such matrices using the combination of speaker-independent and adaptation data, given by (11). HMM parameters for the adapted front end, denoted by $\mathbf{a}^{(\mathrm{Mat + MAP})}$, are then estimated using (7) on training data (11).

4. THE DATABASE AND EXPERIMENTAL FRAMEWORK

To investigate automatic speechreading performance for the speech impaired, we have collected audio-visual speech data of a single speech impaired male subject with profound hearing loss during the period of normal language acquisition. We are interested in audio-visual ASR performance for this subject on both large- and small-vocabulary tasks, therefore two sets of data have been collected (see also Table 1): Approximately 22 minutes (100 utterances) of continuous read speech using ViaVoice$^{\mathrm{TM}}$ dictation scripts (LVCSR task) and about 14 minutes of connected digits strings (140 utterances of 7 to 10 digits each - DIGITS task). Close to half of the collected data have been set aside for testing.

As discussed in the introduction, the amount of collected data is not adequate to train speaker-dependent HMMs. We have therefore chosen to adapt previously trained, speaker-independent HMMs to the impaired data. Such HMMs have been trained using the EM algorithm (see (7)) on the IBM ViaVoice$^{\mathrm{TM}}$ audio-visual database [8], which contains continuous read normal speech with a 10.5 K word vocabulary. In this paper, approximately 35 hours of this database, containing speech from 239 subjects, have been used for HMM training. The resulting HMMs have been adapted to both LVCSR and DIGITS speech impaired tasks, using either the speech impaired LVCSR and DIGITS adaptation sets alone, or jointly. The adapted HMM performances are computed on the speech impaired LVCSR and DIGITS test sets, with a number of adaptation techniques evaluated. Comparisons of the adapted HMMs to normal condition ASR are also performed. For fairness, normal speech recognition results are reported after per subject MLLR adaptation of the speaker-independent LVCSR trained HMM, or, in the case of the DIGITS task, of the *multi-speaker* HMM, trained on a recently collected 50-subject, connected digits database (see Table 1).

All audio-visual data contain full-face frontal video, captured in color, at a size of 704×480 pixels, interlaced, at a rate of 30 Hz (60 fields per second are available at a resolution of 240 lines), and MPEG2 encoded at a 50:1 compression ratio. Wideband audio is synchronously collected with the video at a rate of 16 kHz in an office environment at a 19.5 dB *signal-to-noise ratio* (SNR).

5. EXPERIMENTS

We first consider a number of speaker adaptation techniques, discussed in Section 3, for adapting speaker-independent audio-, visual-only, and audio-visual HMMs trained on the normal speech LVCSR audio-visual data (see Table 1). For each method, the resulting *word error rate* (WER), %, on the speech impaired LVCSR and DIGITS test sets is reported in Table 2. Clearly, the mismatch between the normal and impaired speech data is dramatic, as the "Unadapted" table entries demonstrate. Indeed, the audio-visual WER in the LVCSR task reaches[1] 106.0% (such large numbers occur due to word insertions), whereas the audio-visual WER

---

[1] All impaired speech LVCSR results are reported based on decoding with the 537 word test set vocabulary, unless stated otherwise.

	Method	AU	VI	AV
	Unadapted	116.022	136.359	106.014
L	MLLR (LVCSR-only)	52.266	109.834	42.652
V	MLLR	52.044	110.166	42.873
C	MAP	52.376	101.215	44.199
S	MAP+MLLR	47.624	95.027	41.216
R	Mat+MAP	52.928	98.674	46.519
	Mat+MAP+MLLR	50.055	93.812	41.657
	Unadapted	52.381	48.016	24.801
D	MLLR (DIGITS-only)	5.159	14.881	2.182
I	MLLR	3.770	16.667	0.992
G	MAP	3.373	12.103	1.190
I	MAP+MLLR	2.381	10.516	0.992
T	Mat+MAP	3.968	8.730	1.190
S	Mat+MAP+MLLR	2.381	8.531	0.992

Table 2: Audio- (AU), visual-only (VI), and audio-visual (AV) word error rate, %, on the continuous speech (LVCSR) test set (*upper* table) and on the connected digits (DIGITS) test set (*lower* table) of the speech impaired data using unadapted HMMs (trained in normal speech), as well as a number of HMM adaptation methods. All HMMs are adapted on the joint speech impaired LVCSR and DIGITS adaptation sets, unless stated otherwise. For the continuous speech results, decoding using the test set vocabulary of 537 words is reported.

in the DIGITS task is 24.8% (in comparison, the normal speech, per subject adapted audio-visual LVCSR WER is 10.2%, and the audio-visual DIGITS WER is only 0.55%, computed on the test sets of Table 1).

Subsequently, we apply MLLR HMM adaptation (see (10)) using the speech impaired LVCSR and/or DIGITS adaptation tests. Audio-, visual-only, and audio-visual performances improve dramatically, as demonstrated in Table 2. It is interesting to note that adaptation on joint LVCSR and DIGITS data does help the audio-only performance, however hurts the visual-only results, possibly due to the fact that visual features are less robust to data variability (such as lighting and pose) than audio features. As a result, audio-visual performance does not improve consistently, when comparing within-set (LVCSR or DIGITS) and across-sets (joint LVCSR and DIGITS) MLLR adaptation. Nevertheless, in all remaining experiments, the speaker-independent LVCSR HMMs are adapted on the joint adaptation set. Next, we consider MAP HMM adaptation (see Section 3.2). Due to the rather large adaptation set, MAP performs similarly well to MLLR. Applying MLLR after MAP improves results, and it reduces the audio-visual WER to 41.2% and 0.99% for the LVCSR and DIGITS tasks, respectively.[2] This cor-

---

[2]Notice that the LVCSR MAP+MLLR audio-, visual-only, and audio-visual results become 64.6%, 102.4%, and 58.5%, respectively, when using
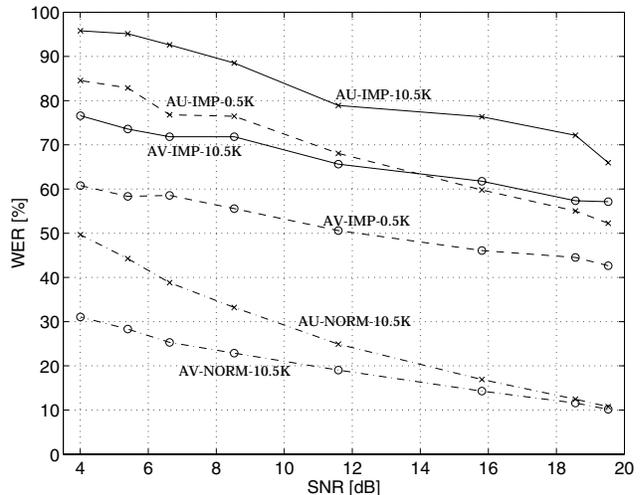

Figure 2: Comparison of audio-only (AU) and audio-visual (AV) automatic speech recognition of impaired (IMP) and of normal speech (NORM), after MLLR adaptation, in the continuous, read speech (LVCSR) domain. The word error rate (WER), %, on the corresponding test set (see Table 1) is depicted as a function of the signal-to-noise ratio (SNR) present in the audio channel. *Solid lines:* WER for the speech impaired subject, decoded using a 10.5 K word vocabulary. *Dash lines:* Speech impaired subject WER, decoded using the test set 0.5 K vocabulary. *Dash-dot lines:* Normal speech, per subject adapted WER, decoded using the 10.5 K word vocabulary.

responds to a 61% and 96% relative WER reduction over the audio-visual unadapted results, and to a 13% and 58% relative WER reduction over the audio-only MAP+MLLR adapted results, for the two recognition tasks, respectively. Clearly, therefore, the visual modality dramatically benefits the automatic recognition of impaired speech.

We also apply front end adaptation (Section 3.3), possibly followed by MLLR adaptation, with the results depicted in the Mat+MAP(+MLLR) entries of Table 2. Although visual-only recognition improves, the audio-only recognition results fail to do so. As a consequence, audio-visual ASR degrades, possibly also due to the fact that, in this experiment, audio-visual matrix adaptation is only applied to the second stage of LDA/MLLT. It is worth mentioning that the resulting speaker-adapted visual-only impaired speech recognition WER of 93.8% in the LVCSR task (this becomes 97.2%, when using the 10.5 K word vocabulary) is worse than the normal condition visual-only 89.2% WER, however in the DIGITS task, the 8.53% impaired speech visual-only WER is significantly better than the normal condition 16.77% WER.

In addition to the adaptation experiments using the original database audio (at 19.5 dB SNR), we consider audio-only
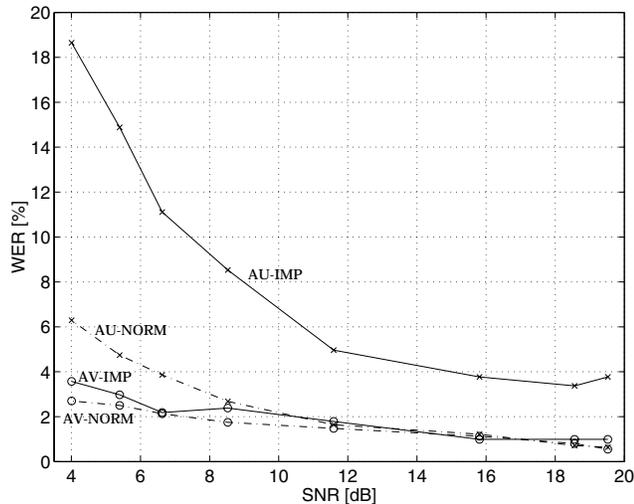
---

the 10.5 K word decoding vocabulary.

Figure 3: Audio-only and audio-visual ASR performance on impaired (*solid lines*) and on normal (*dash-dot lines*) speech for the connected digits (DIGITS) task, depicted in WER as a function of the audio channel SNR. Per subject MLLR adaptation is used.

and audio-visual recognition for the LVCSR and DIGITS domains, when the audio channel is artificially corrupted by additive "*babble*" speech (cafeteria like) noise at a number of SNR values. Both normal and impaired speech recognition are compared, with all HMMs and LDA/MLLT matrices trained at the *matched* condition and adapted using MLLR on joint LVCSR/DIGITS impaired data at the same SNR level. The results are depicted in Figures 2 and 3, for the LVCSR and DIGITS tasks, respectively. Interestingly, for the DIGITS task, speech impaired audio-visual ASR performance is very close to audio-visual ASR of normal speech. This is due to the superior visual-only recognition of the speech impaired DIGITS. However, in the LVCSR domain, the visual channel does not provide sufficient speech information, and, as a result, speech impaired audio-visual ASR is significantly worse than LVCSR of subjects with normal speech.

6. CONCLUSIONS AND FUTURE WORK

We investigated the use of machine speechreading to improve automatic speech recognition of the speech impaired. We considered both a small- (connected digits) and a large-vocabulary recognition task, and we applied HMM and front end adaptation techniques to improve audio-visual impaired speech recognition. By using a combination of MAP and MLLR adaptation, we achieved a 58% and 13% relative WER reduction over the equivalent audio-only system for the small- and large-vocabulary tasks, respectively. For the former task, audio-visual recognition of impaired speech reached the recognition performance of normal audio-visual speech, over a wide range of audio channel SNRs.

A word of caution is warranted about our conclusions, due to the fact that data from a single speech impaired subject have been collected and used in our experiments. We plan to investigate the generalization of these results to a larger speech impaired population in the near future.

7. REFERENCES

[1] Campbell, R., Dodd, B., and Burnham, D. eds., *Hearing by Eye II*. Psychology Press Ltd. Publishers, Hove, 1998.

[2] Stork, D.G. and Hennecke, M.E. eds., *Speechreading by Humans and Machines*. Springer, Berlin, 1996.

[3] Sumby, W.H. and Pollack, I., "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. America*, vol. 26, pp. 212–215, 1954.

[4] McGurk, H. and MacDonald, J.W., "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] Hennecke, M.E., Stork, D.G., and Prasad, K.V., "Visionary speech: Looking ahead to practical speechreading systems," in [2], pp. 331–349, 1996.

[6] Dupont, S. and Luettin, J., "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, vol. 2, pp. 141–151, 2000.

[7] Chen, T., "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Process. Mag.*, vol. 18, pp. 9–21, 2001.

[8] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J., "Audio-visual speech recognition," *Final Workshop 2000 Report*, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, 2000 (http://www.clsp.jhu.edu/ws2000/final_reports/avsr/).

[9] Potamianos, G., Luettin, J., and Neti, C., "Hierarchical discriminant features for audio-visual LVCSR," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 1, 2001.

[10] Bernstein, L.E., Demorest, M.E., and Tucker, P.E., "What makes a good speechreader? First you have to find one," in [1], pp. 211–227, 1998.

[11] Marschark, M., LePoutre, D., and Bement, L., "Mouth movement and signed communication," in [1], pp. 245–266, 1998.

[12] Rabiner, L. and Juang, B.-H., *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.

[13] Gauvain, J.-L. and Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, pp. 291–298, 1994.

[14] Neumeyer, L., Sankar, A., and Digalakis, V., "A comparative study of speaker adaptation techniques," *Proc. EUROSPEECH*, pp. 1127–1130, 1995.

[15] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Lang.*, vol. 9, pp. 171–185, 1995.

[16] Potamianos, G. and Potamianos, A., "Speaker adaptation for audio-visual speech recognition," *Proc. EUROSPEECH*, vol. 3, pp. 1291–1294, 1999.

[17] Senior, A.W., "Face and feature finding for a face recognition system," *Proc. Int. Conf. Audio Video-based Biometr. Person Authent.*, pp. 154–159, 1999.