USING THE MULTI-STREAM APPROACH FOR CONTINUOUS AUDIO-VISUAL SPEECH RECOGNITION: EXPERIMENTS ON THE M2VTS DATABASE

Stéphane Dupont[1,†,‡] *& Juergen Luettin*[‡]

† Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez, B-7000 Mons, Belgium
Email: dupont@tcts.fpms.ac.be

‡ Institut Dalle Molle d'Intelligence Artificielle Perceptive
BOX 592, rue du Simplon 4, 1920 Martigny, Switzerland
Email: luettin@idiap.ch

ABSTRACT

The Multi-Stream automatic speech recognition approach was investigated in this work as a framework for Audio-Visual data fusion and speech recognition. This method presents many potential advantages for such a task. It particularly allows for synchronous decoding of continuous speech while still allowing for some asynchrony of the visual and acoustic information streams. First, the Multi-Stream formalism is briefly recalled. Then, on top of the Multi-Stream motivations, experiments on the M2VTS multimodal database are presented and discussed. To our knowledge, these are the first experiments addressing multi-speaker continuous Audio-Visual Speech Recognition (AVSR). It is shown that the Multi-Stream approach can yield improved Audio-Visual speech recognition performance when the acoustic signal is corrupted by noise as well as for clean speech.

1. INTRODUCTION

The Multi-Stream approach used in this work is a principled way for merging different sources of information. In this approach, it is assumed that the speech signal is described in terms of multiple input streams, each stream representing a different characteristic of the input signal. If the streams are supposed to be entirely synchronous, they may be accommodated simply. However, it is often the case that the streams are not synchronous, that they do not even have the same frame rate and it might be useful to define models that do not have the same topology. The Multi-Stream approach discussed in [2] allows to deal with this. In this framework, the input streams are processed independently of each other up to certain anchor points where they have to synchronize and recombine their partial segment-based likelihoods. While the phonological level of recombination has to be defined a priori, the optimal temporal anchor points are obtained automatically during recognition.

The subband-based speech recognition approach, a particular case of Multi-Stream, was shown on several databases to yield significantly better noise robustness [3, 10] compared to standard approaches. The general idea of this subband-based approach is to split the whole frequency band (represented in terms of critical bands) into a few subbands on which different recognizers are independently applied and then recombined at a certain speech unit level to yield global scores and a global recognition decision. This subband-based approach has many other motivations, including the possibility to better accommodate the possible asynchrony between different components of the speech spectrum [12].

Another application that was investigated recently is the possibility to incorporate multiple time resolutions as part of a structure with multiple length units, such as phone and syllable. In the same framework, it is indeed possible to define subword models composed of several cooperative HMM models focusing on different dynamic properties of the speech signal. Preliminary results were presented in [5].

The feature that will be investigated here is the possibility to combine several information sources. The Multi-Stream formalism and decoding scheme will indeed be used as framework for an Audio-Visual continuous speech recognition system.

2. AUTOMATIC AVSR AND THE MULTI-STREAM APPROACH

Speech-reading as well as integration of auditory and visual parameters for speech recognition has gained interest in the scientific community these past few years [7, 8, 11, 9]. This is probably because Audio-Visual integration offers many potential advantages for automatic speech recognition systems. Several studies have indeed shown that the use of lip movement information, in addition to the acoustics, can significantly improve the recognition performance in the case of speech corrupted by acoustic noise. Moreover, it is acknowledged that the acoustics and the lip movements carry complementary information. For instance, discriminating between the phonemes /t/ and /p/ can be easier with the visual information than with the acoustic information. A more extensive insight into the problem can be found in other publications [9].

This work was particularly motivated by the fact that the Multi-Stream formalism, introduced earlier as framework for subband-based speech recognition [3] and then used for multiscale-based speech recognition [5], could be an efficient approach for continuous Audio-Visual speech recognition. Other contributors, cited in the previous paragraph, have essentially addressed the problem of isolated word recognition. The proposed approaches were based on a recombination of likelihoods from the visual and acoustic streams at the end of the uttered word, or on the feature combination at the frame level. Most of these contributions were mainly focused on finding an appropriate automatic weighting scheme so as to guarantee good performance in a wide range of acoustic signal-to-noise ratios.

Compared to isolated word recognition, the problem of continuous speech recognition is more tricky as we do not

[1] Supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture), Belgium

want to wait until the end of the spoken utterance before recombining the streams. Indeed, this introduces a time delay and this also requires to generate N-best hypothesis lists for the two streams. Indeed, one can only recombine the scores from identical hypothesis. As the best hypothesis for the acoustic stream is not necessarily the same as the best hypothesis for the visual stream, N-best lists are required. Identical hypothesis must then be matched to recombine the scores from the two streams. An alternative approach would be to generate an N-best list for one of the two streams, to compute the score of these best hypothesis for the other stream, and finally to recombine the scores. The Multi-Stream approach does not require to use such an N-best scheme and is an interesting candidate for multimodal continuous speech recognition as it allows for:

- **synchronous multimodal continuous speech recognition**: by using the*HMM recombination* or the *two-level* decoding schemes [2].
- **asynchrony of the visual and acoustic streams**: The Multi-Stream approach can force the two modalities to be synchronous where synchrony is required and can allow for asynchrony of the visual and acoustic stream where asynchrony might take place. The level of required synchrony might be chosen heuristically, as was done in the experiments presented in this paper, or might be learned from training data (see Section 4).
- **specific audio and video word or sub-word topologies**: The Multi-Stream model is composed of parallel models which do not necessarily have the same topologies.

Tomlinson and al. [11] already addressed the issues of visual and acoustic components *asynchrony* and *continuous Audio-Visual speech recognition*. The technique was based on HMM decomposition. Under the independence assumption, composite models were defined from independently trained audio and visual models. Although our work is related with [11], it allows to consider different recombination formalisms and enables the decoding of continuous speech. Moreover, the scope of asynchrony between the two streams was here extended from the phone level to the word level.

The next section presents speech recognition experiments done on the multimodal M2VTS database. In addition to using the novel Multi-Stream scheme, this work is one of the first that addresses multi-speaker continuous Audio-Visual speech recognition.

3. EXPERIMENTS ON THE M2VTS MULTIMODAL FACE DATABASE

The M2VTS database was collected as part of the M2VTS project granted by the European ACTS program. The primary goal of the M2VTS project (Multi Modal Verification for Teleservices and Security applications) was to address the problem of secured access to buildings by using multimodal identification/verification methods. The database is thus made of synchronized images and speech as well as sequences allowing to access multiple views of a face.

The part of the database we have been using in this work consists of 37 different persons, each pronouncing 5 times the sequence of digits from '0' to '9' in French. This is the only part of the database which can be used for multimodal speech recognition, the rest of the database consisting of people rotating their head. The video sequences consist in 286*360 pixel color images with a 25 Hz frame rate and the audio track was recorded at a 48 kHz sampling frequency and 16 bit PCM coding. Further information can be found in [1].

3.1. Database Partitioning

Although M2VTS is the largest database of its type, it is still relatively small compared to audio databases used in the field of speech recognition. To increase the significance level of our experiments, we used a jack-knife approach. Five different cuts of the database were used. Each cut consisted of:

- 3 pronunciations from the 37 speakers as training set.
- 1 pronunciation from the 37 speakers as development set. It was used to optimize parameters such as weighting coefficients between audio and video streams.
- 1 pronunciation from the 37 speakers as test set.

This procedure allowed to use the whole database as test set (185 utterances) by developing five independent speech recognition systems for each of the compared approaches. These systems could be qualified as multi-speaker (but speaker dependent) continuous digits recognition systems. We note here that the digit sequence to be recognized is always the same (digits from '0' to '9'). This somewhat simplifies the task of the speech recognition system which always "see" the pronounced words in the same context.

Systems were first developed to recognize the pronounced digit sequences using the information conveyed by the audio stream only or by the video stream only.

3.2. Audio-based Speech Recognition

The audio stream was first downsampled to 8 kHz. We used PLP parameters [6], computed every 10 ms on 30 ms sample frames. The complete feature vectors consisted of 25 parameters: 12 PLP coefficients, 12 ΔPLP coefficients and the Δenergy.

We used left-right digit HMM models with between 3 and 9 independent states, depending on the digit mean duration. This yielded a total of 52 states. The digit sequences were first segmented into digits using standard Viterbi alignment with a system trained on the SWISS-FRENCH POLYPHONE database [4]. Each digit was then linearly segmented according to the number of states of the corresponding HMM model. This segmentation was used to train HMM states, whose emmission probability was modelled by a mixture of two multi-dimensional Gaussian distributions with diagonal covariance matrices, yielding a total of 5200 parameters. Iterative Viterbi alignment and reestimation of the model parameters was also performed.

System training and test were performed according to the database partitioning described earlier. Results are summarized in Figure 3 for clean speech as well as for speech corrupted by additive white noise with different signal-to-noise ratios. As can be observed, recognition performance is severely affected by additive noise, even at such moderate noise levels.

3.3. Video-based Speech Recognition

In this case, geometric features and grey-level features from the mouth region were used, assuming that they carry relevant linguistic information. An appearance based model of the articulators is learned from example images and is used to locate, track and recover visual speech features [7]. The method decomposes the lip shape and the grey-level intensities in the mouth region into a weighted sum of basis shapes (inner and outer lip contour) and basis intensities, respectively, using the Karhunen-Loeve expansion. These features, obtained by lip tracking, were normalized with respect to the mouth center, orientation, and width. The 12 most relevant shape features and the 12 most relevant intensity features together with their temporal difference parameters, yielding 48 parameters, were used for the HMM based speech recognition system.

We used the same HMM topologies and the same initial segmentation as for the previously described acoustic-based recognition system. In this case, the emission probabilities of the HMM states were modeled by a single mixture multidimensional Gaussian distributions with diagonal covariance matrices.

The mean error rate for the five database cuts defined earlier was 43.9%.

Since the visual signal only provides partial information, the error rate for the video-based system was considerably higher than for the audio-based system This is mainly due to the high visual similarity of certain digits like “quatre”, “cinq”, “six”, and “sept”. About half of the errors were due to substitutions of these highly confusable digits and the other half were caused by deletion errors.

3.4. Multimodal Speech Recognition

Audio-Visual speech recognition was experimentally investigated and 2 kinds of model topologies were compared. These were based on the HMM word topologies already used in the two previous sections. The differences between the models lay in the possible asynchrony of the visual stream with respect to the acoustic stream. In the experiments that were carried out, the word topologies were the same for the two modalities. Let's recall however that the Multi-Stream approach would allow to use specific audio and video HMM topologies.

The first model (Model 1) did not allow for any desynchronization between the two streams. It corresponds to a Multi-Stream model with recombination at the state level and allows to use fusion criteria that weight the two streams differently according to their respective reliability.

The second model (Model 2) was a Multi-Stream model with recombination of the streams at the word level. This model thus allows the dynamic programming paths to be independent from the beginning up to the end of the words. This relaxes the assumption of piecewise stationarity by allowing the stationarity of the two streams to occur on different time regions, while still forcing the modalities to resynchronize at word boundaries. This also accounts for the possible asynchrony of the streams inherent to the speech production mechanism. Indeed, lip movements and changes in the vocal tract shape are independent up to a point.

Model 2 also allows the transition from silence to speech and from speech to silence to occur at different time instants for the two streams [2]. Indeed, it seems likely that lip movement can occur before and after sound production and conversely. Figure 1 shows in parallel a speech spectrogram as well as the evolution of the first visual shape

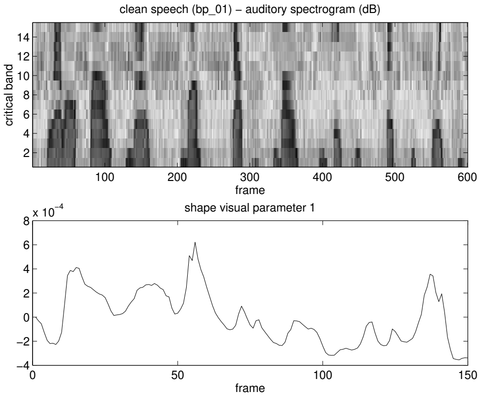

Figure 1. Auditory spectrogram (evolution of the critical band energies) and evolution of the first visual shape parameter for one portion (’0’ to ’8’) of an M2VTS utterance.

parameter, mainly representing the changes in the position of the lower lip contour [7]. It can clearly be seen that the two signals are partially in synchrony and partially asynchronous. Ideally, we would like to have a model which forces the streams to be synchronous where synchrony occurs and asynchronous where the signals are typically in asynchrony. Model 2 for a particular vocabulary word is presented in Figure 2. The model is composed of two parallel HMMs, associated with the two modalities. The recombination state ($\bigotimes$) is not a regular HMM state since it will be responsible for recombining probabilities (or likelihoods) accumulated over the same temporal segment for the acoustic and visual modalities.

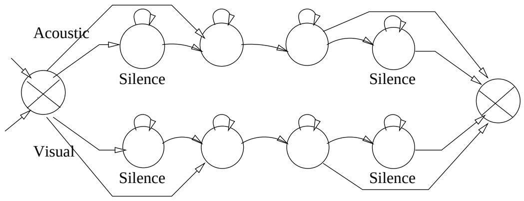

Figure 2. Multi-stream model for Audio-Visual speech recognition with optional silence states (Model 2).

We used the same parameterization schemes as in the two previous sections. However, as the visual frame rate (25 Hz) is a quarter of the acoustic frame rate, visual vectors were artificial added at the probability level (by copying frames), so that acoustic and video stay synchronous, simplifying the data fusion implementation.

In this work, recombination of the independent likelihoods was done linearly, by multiplying segment (sub-units) likelihoods from the two streams, thus assuming conditional independence of the visual and acoustic streams. This was done according to:

$$p(X|M) = p(X_{acou}|M_{acou})^{w}.p(X_{vis}|M_{vis})^{(1-w)}, \qquad (1)$$

where $p(X_{acou}|M_{acou})$ represents the likelihood of the time limited acoustic information stream given the acoustic part of model M (lexical sub-unit model), $p(X_{vis}|M_{vis})$ is the corresponding likelihood for the visual stream and w is a parameter allowing to weight the two streams according to their respective reliability. It was optimized on the development set. If Model 1 is used, this simply boils down to multiplying local likelihoods. For the other model (Model 2), this multiplication has to be

[2]'Visual silence' could be defined as a portion of the visual signal that doesn't carry any relevant linguistic information.

performed at the word boundaries. This was done in a synchronous way using the HMM recombination algorithm [2], an adaptation of the HMM decomposition algorithm [13].

System	Video	Audio	Audio-Visual
Error rate	43.9%	3.4%	2.6%

Table 1. Word error rate of audio-, video- and Audio-Visual-based (MODEL 2) speech recognition systems on clean speech.

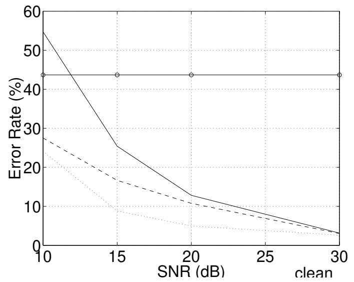

Figure 3. Word error rate of audio-, video- and Audio-Visual-based speech recognition systems at different acoustic SNR levels. This graph presents the results obtained after embedded training of the 2 kind of models and of the acoustic-only model (training on clean speech only). The solid line is for the acoustic system, the dashed line is for MODEL1 and the dotted line is for MODEL 2. The horizontal line represents the performance of the visual-only system.

In these experiments, the optimal recombination weight was optimized on the development set for each of the test conditions. Consequently, these results do not represent what could be achieved with a practical system. In practice, one should design an automatic way of estimating the recombination parameter. One way could be to define a mapping between this parameter and an SNR estimate. From our experiments, it can be seen that the optimal weight is related almost linearly to the SNR ratio and can easily be estimated from it.

Results are summarized in Figure 3 for different levels of noise degradation. In the case of clean speech, using visual information, in addition to the acoustics, does not yield significant performance improvement (see Table 1). The confidence level of the hypothesis test was 0.95. In the case of speech corrupted with additive stationary Gaussian white noise, significant performance improvement can be obtained by using the visual stream as an additional information source. The results also clearly show that we can get a significant performance improvement with MODEL2 compared to MODEL1 by allowing the acoustic and visual decoding paths to be in asynchronous.

It should be noted, however, that decoding with Multi-Stream models is somewhat more complex than decoding using models that constrain the streams to be synchronous (Multi-Stream with state-level recombination - MODEL 1). Computational requirements significantly increase but stay within an order of magnitude above the classical model. In the next Section, we will propose a method that would allow to reduce the computational load.

4. CONCLUSIONS

We have presented a framework for the fusion of acoustic and visual information in an Audio-Visual speech recognition system based on the Multi-Stream approach. Several significant advances have been reported in this paper. Firstly, the method enables synchronous Audio-Visual decoding of continuous speech and we have presented one of the first continuous AV speech recognition experiments. Secondly, it allows for asynchronous modeling of the two streams, which is inherent in the acoustic and visual speech signal and which has been shown to lead to more accurate modeling and to improved performance. Thirdly, the approach allows to design specific Audio-Visual word or sub-word topologies. This also includes the modeling of possible monomodal or multimodal "silence" states.

ACKNOWLEDGEMENTS

We would like to thank Hervé Bourlard from IDIAP for his support and for many useful discussions.

REFERENCES

[1] "The M2VTS Database WWW site." `http://www.tele.ucl.ac.be/M2VTS/`.

[2] H. Bourlard, S. Dupont, and C. Ris, "Multi-stream speech recognition," Tech. Rep. IDIAP-RR 96-07, IDIAP, Martigny, Switzerland, 1996.

[3] H. Bourlard and S. Dupont, "Sub-band-based speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, (Munich), Apr. 1997.

[4] A. Constantinescu, O. Bornet, G. Caloz, and G. Chollet, "Validating different flexible vocabulary approaches on the swiss french polyphone and polyvar databases," in *Proc. of the Intl. Conf. on Spoken Language Processing.*, (Philadelphia, PA), September 1996.

[5] S. Dupont and H. Bourlard, "Using multiple time scales in a multi-stream speech recognition system," in *Proc. of EUROSPEECH'97.*, (Rhodes, Greece), Sept. 1997.

[6] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, Apr. 1990.

[7] J. Luettin and N. Thacker, "Speechreading using probabilistic models," *Computer Vision and Image Understanding*, vol. 65, pp. 163–178, February 1997.

[8] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Lawrence Erlbaum Associates, 1987.

[9] D. G. Stork and M. E. Hennecke, eds., *Speechreading by Humans and Machines: Models, Systems and Applications.* Berlin: NATO ASI Series F, Computer and Systems Sciences, Springer-Verlag, 1996.

[10] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. of ICASSP'97*, (Munich), pp. 1255–1258, 1997.

[11] M. Tomlinson, M. Russel, and N. Brooke, "Integrating audio and visual information to provide highly robust speech recognition," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, 1996.

[12] M. Tomlinson, M. Russell, R. Moore, A. Buckland, and M. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition," in *Proc. of ICASSP'97*, (Munich), pp. 1247–1250, 1997.

[13] A. Varga and R. Moore, "Hidden markov model decomposition of speech and noise," in *Proc. IEEE Internat. Conf. Acoust. Speech and Signal Process.*, pp. 845–848, 1990.