Recapitulation on Audio-Visual features for Speech Recognition

Anusha Kistapuram\(^1\) & Srinivas Jammikuntla\(^2\)
Sreenidhi Institute of Science and Technology

Abstract--Speech recognition using different modalities had paved the way for tremendous applications and novel advancements in linguistics, emotion recognition, human computer interface, and associated disciplines. My focus is on bimodal recognition process using modalities from both visual front and audio. I will review on diverse prone bimodal databases, browsing from modest to broad terminology understanding exercise. Few inquiries exhibit that visual modality outperforms speech recognition over all circumstances and confronting surroundings. However, collective modelling of audio and visual features is a great challenge to aid for decision making in terms of recognition. This paper aims for comprehensive and analysis of state of art schemes using audio visual features and their significance.

Keywords—Speech recognition; bimodal; visual; audio; features

I. Introduction

Understanding human speech implicates investigation of articulated phonics and incorporates superior insight into grammar, semantics and pragmatics [1]. In order to serve hearing impaired, visual modalities play a major role to an extent of naive speech perception. Speech recognition is considered as an indispensable chunk of forthcoming human-computer hook up, that are anticipated to adopt speech, among alternative modes, to accomplish natural, extensive and ever present computing. In noisy environments, intelligibility of speech improves with lip movements rather than relying on traditional audio cues. [2-3.

The main advantage of the visual signal is its independence to the acoustic signal [4]. Phonemes that is most difficult to perceive in the presence of noise are easier to distinguish visually and vice versa. The visual signal contains that kind of information that is acoustically most sensitive to noise. Studies have also shown that visual information leads to more accurate speech perception even in noise-free environments [5]. The strong influence of visual speech cues on human speech perception is demonstrated by the McGurk effect [6] in which, for example, a person hearing an audio recording of /baba/ and seeing the synchronised video of a person saying /dada/ often resulted in perceiving /gaga/. [1]There are three key reasons why vision benefits human speech perception [7]: It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the place of articulation. The latter is due to the partial or full visibility of articulators, such as the tongue, teeth, and lips. Place of articulation information can help disambiguate, for example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) from the nasal alveolar /n/ [5]. All three pairs are highly confusable on basis of acoustics alone. In addition, jaw and lower face muscle movement is correlated to the produced acoustics [3–5], and its visibility has been demonstrated to enhance human speech perception [7].

The above facts have motivated significant interest in automatic recognition of visual speech, formally known as automatic lip reading or speech reading [5]. Work in this field aims at improving ASR by exploiting the visual modality of the speaker’s mouth region in addition to the traditional audio modality, leading to audio-visual automatic speech recognition (AV-ASR) systems. Compared to audio-only speech recognition, AV-ASR introduces new and challenging tasks, First, in addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker’s face. This requires robust face detection, as well as location estimation and tracking of the speaker’s mouth or lips, followed by extraction of suitable visual features. In contrast to audio-only, recognizers, there are now two streams of features available for recognition, one for each modality. The combination of the audio and visual streams should ensure that the resulting system performance is better than the better of the two single modality recognizers, and hopefully, significantly outperform it. Both issues, namely the visual front end design and audio-visual fusion, constitute difficult problems [6], and they have generated much research work by the scientific community.
The first automatic speech reading system was reported in 1984 by Peterman [4]. Given the video of the speaker’s face, and by using simple image thresholding, he was able to extract binary (black and white) mouth images, and subsequently, mouth height, width, perimeter, and area, as visual speech features. He then developed a visual-only recognizer based on dynamic time warping [3] to restore the best two choices of the output of the baseline audio-only system. His method improved ASR for a single-speaker, isolated word recognition task on a 100-word vocabulary that included digits and letters. Since then, over a hundred articles have concentrated on AV-ASR, with the vast majority appearing during the last decade. The reported systems differ in three main aspects [1]: The visual front end design, the audio-visual integration strategy, and the speech recognition method used. Unfortunately, the diverse algorithms suggested in the literature are very difficult to compare, as they are rarely tested on a common audio-visual database. Nevertheless, the majority of systems outer-form audio-only ASR over a wide range of conditions. Such improvements have been typically demonstrated on databases of small duration, and, in most cases, limited to a very small number of speakers (mostly less than ten, and often single-subject) and to small vocabulary tasks [1], [2]. Common tasks typically include recognition of non-sense words [22], [3], isolated words [7], connected digits [3], [2], letters [4], or of closed-set sentences [3], mostly in English, but also in French [2], [4], [5], German [6], [7], and Japanese [8], among others. Recently however, significant improvements have also been demonstrated for large vocabulary continuous speech recognition (LVCSR) [3], as well as cases of speech degraded due to speech impairment [4] or Lombard effects [2]. These facts, when coupled with the diminishing cost of quality video capturing systems, make automatic speech reading tractable for achieving robust ASR in certain scenarios and tasks [1].

In more detail, Section-II of the paper concentrates on the visual front end, first summarizing relevant work in the literature, and subsequently discussing its three main blocks in our system. It also reports recent improvements in the extraction and normalization of the visual region of interest. Section III presents issues in visual speech modelling that are relevant to audio-visual fusion. Section-IV is devoted to an overview of audio-visual fusion, considering three classes of algorithms, i.e., feature, decision, and hybrid fusion. In particular, it introduces a novel technique within the last category, and also discusses the issue of audio-visual asynchrony modelling. Section-V concentrates on a very important aspect of decision fusion based AV-ASR, namely modelling the reliability of the audio and visual stream information. A number of local stream reliability indicators are considered, and a function that maps their values to appropriate decision fusion parameters is introduced. Section-VI reports correlative findings on them. Finally, Section-VII concludes the paper with a summary and a brief discussion on the Current state and open problems in AV-modelling.

II. Visual Speech Feature Extraction

Facial feature extraction is a difficult problem due to large appearance differences across persons and due to appearance variability during speech production. Different illumination conditions and different face positions cause further difficulties in image analysis. For a real-world application, whether it is in a car, an office or a factory, the system should be able to deal with these kinds of image variability. Illustration is shown in Fig1.

The main approaches for extracting visual speech information from image sequences can be grouped into the following approaches:

1) image-based;
2) visual-motion-based;
3) geometric-feature-based; and
4) model-based.

In the image-based approach [1]-[3], he grey-level image containing the mouth is either used directly or after some image transform as feature vector whereas the visual-motion-based method [4] assumes that visual motion during speech production contains relevant speech information. Geometric-feature-based techniques [5], on the other hand, assume that certain measures such as the height or width of the mouth opening are important features. Finally, in the model-based approach [6], a model of the visible speech articulators, usually the lip contours, is built and its configuration is described by a small set of parameters. The advantage of the latter approach is that important features can be represented in a low-dimensional space and can often be made invariant to image transforms like translation, scaling, rotation and lighting. A disadvantage is that the particular model used may not consider all relevant speech information. The main difficulty in the model-based approach is the definition of the model and the development of image search procedures that accurately find the correspondence between the model and the image.
III. Visual Speech Modelling for ASR

Once features become available from the visual front end, one can proceed with automatic recognition of the extracted acoustic features for audio-visual ASR (see also Fig.1). The first scenario is primarily useful in benchmarking the performance of visual feature extraction algorithms, with visual-only ASR results typically reported on small vocabulary tasks [2]. Visual speech modelling is required in this process, its two central aspects being the choice of speech classes that are assumed to generate the observed features, and the statistical modelling of this generation process. Both issues are important, as they are also embedded into the design of audio-visual fusion discussed next.

Table 1 illustrates a case study of phoneme to visual mapping.

<table>
<thead>
<tr>
<th>Visual class</th>
<th>Phonemes in cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silence</td>
<td>/h/, /h/</td>
</tr>
<tr>
<td>Lip-rounding</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Dorsal vowels</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Alveolar-semivowels</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Alveolar-fricatives</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Palato-alveolar</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Bilabial</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Dental</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
<tr>
<td>Labio-dental</td>
<td>/tul/, /tul/, /tul/</td>
</tr>
</tbody>
</table>

Table 1 [39]

IV. Audio Visual Integration for ASR

As already mentioned in Section I, audio-visual integration constitutes a major research topic in AV-ASR, aiming at the combination of the two available speech informative streams into a bimodal classifier with superior performance to both audio- and visual-only recognition. Various information fusion algorithms have been considered for AV-ASR, differing both in their basic design, as well as in the terminology used [1], [2], and [7]. In this paper, we adopt their broad grouping into feature fusion and decision fusion methods. The first are based on training a single classifier (i.e., of the same form as the audio- and visual-only classifiers) on the concatenated vector of audio and visual features, or on any appropriate transformation of it [2]. In contrast, decision fusion algorithms utilize the two single-modality (audio- and visual-only) classifier outputs to recognize audio-visual speech. Typically, this is achieved by linearly combining the class-conditional observation log-likelihoods of the two classifiers into a joint audio-visual classification score, using appropriate weights that capture the reliability of each single-modality classifier, or data stream [7], [3].

V. Stream Reliability Modelling AV-ASR

I now address the issue of stream exponent (weight) estimation, when combining likelihoods in the audio-visual decision and hybrid fusion models. There, such exponents are set to constant stream-dependent values, to be computed for a particular audio-visual environment and database, based on the available training, or more often, held-out data. Due to the form of the emission scores, the stream exponents cannot be obtained by maximum likelihood estimation [3]. Instead, discriminative training techniques are used.

Some of these methods seek to minimize a smooth function of the word minimum classification error (MCE) of the resulting audio-visual model on the data, and employ the generalized probabilistic descent (GPD) algorithm for stream exponent estimation [1], [3]. Other techniques use maximum mutual information (MMI) training as in [4]. A different approach minimizes the frame error rate, by using the maximum entropy criterion alternatively; one can seek to directly minimize the word error rate of the resulting audio-visual ASR system on a held-out data set. In the case of two global exponents, constrained to add to a constant, the problem reduces to one-dimensional optimization of a non-smooth function, and can be solved using simple grid search. In the case where additional streams are present, or when class dependent stream exponents are desired, the problem becomes of higher dimension, and the downhill simplex method can be employed. In general however, class dependency has not been demonstrated to be effective in AV-ASR [4], with the exception of the late integration, discriminative model combination technique.

Therefore, in this paper, class dependency of global stream exponents is not considered. These global exponents are then simply estimated by grid search on a held-out set.
VI. Correlative Findings (Exclusive-Study).

Fig. 2. Audio-only and audio-visual WER, % for the studio-LVCSR database test set using Hilda feature fusion, and two multi-stream HMMs, using a visual-only, or Hilda feature stream. Courtesy: [1]

VII. SUMMARY

I provided a brief overview of the basic techniques for automatic recognition of audio-visual speech, proposed in the literature over the past twenty years, with particular emphasis in the algorithms used in our speech reading system. The two main issues relevant to the design of audio-visual ASR systems are: First, the visual front end that captures visual speech information and, second, the integration (fusion) of audio and visual features into the automatic speech recognizer used. Both are challenging problems, and significant research effort has been directed towards finding appropriate solutions. The best algorithms resulted in an effective SNR gain of 10dB for connected digital recognition of visually "clean" data [1]. The gains were less, for a large-vocabulary task (8dB), as well as for visually challenging data.

Clearly, over the past twenty years, much progress has been accomplished in capturing and integrating visual speech information into automatic speech recognition. However, the visual modality has yet to become utilized in mainstream ASR systems. This is due to the fact that issues of both practical and research nature remain challenging. On the practical side of things, the high quality of captured visual data, which is necessary for extracting visual speech information capable of enhancing ASR performance, introduces increased cost, storage and computer processing requirements. In addition, the lack of common, large audio-visual corpora that address a wide variety of ASR tasks, conditions, environments, hinders development of audio-visual systems suitable for use in particular applications.

On the research side, the key issues in the design of Audio-visual ASR systems remain open and subject to more investigation. In the visual front end design, for example, face detection, facial feature localization, and face shape tracking, robust to speaker, pose, lighting, and environment variation constitute challenging problems. A comprehensive comparison between face appearance and shape based features for speaker-dependent vs. speaker-independent automatic speech reading is also unavailable. Joint shape and appearance three-dimensional face modelling; used for both tracking and visual feature extraction has not been considered in the literature, although such an approach could possibly lead to the desired robustness and generality of the visual front end. In addition, when combining audio and visual information, a number of issues relevant to decision fusion require further study, such as the optimal level of integrating the audio and visual log-likelihoods and the optimal function for this integration.

REFERENCES