An Audio-Visual Speech Recognition Framework Based on Articulatory Features

Tian Gan, Wolfgang Menzel, Shiqiang Yang

Department of Informatics, University of Hamburg
Department of Computer Science and Technology, Tsinghua University
{gan, menzel}@informatik.uni-hamburg.de
{yangshq}@tsinghua.edu.cn

Abstract

This paper presents an audio-visual speech recognition framework based on articulatory features, which tries to combine the advantages of both areas, and shows a better recognition accuracy compared to a phone-based recognizer. In our approach, we use HMMs to model abstract articulatory classes, which are extracted in parallel from both the speech signal and the video frames. The N-best outputs of these independent classifiers are combined to decide on the best articulatory feature tuples. By mapping these tuples to phones, a phone stream can be generated. A lexical search finally maps this phone stream to meaningful word transcriptions. We demonstrate the potential of our approach by a preliminary experiment on the GRID database, which contains continuous English voice commands for a small vocabulary task.

Index Terms: audio-visual speech recognition, articulatory features, N-best decision schema

1. Introduction

After almost 30 years of research in Automated Speech Recognition (ASR), scientists have covered applications ranging from speaker dependent, isolated word recognition, to speaker independent, large vocabulary, continuous speech recognition [1] and [2]. The technology has reached a level of performance which seems difficult to be improved further, if only acoustic evidence is considered. On the other hand, most of the currently available systems require proper acoustic conditions, including a quiet environment, good quality microphones, a suitable distance to the microphone, etc. There is a clear necessity to overcome these limitations by including additional speech-related information into the decision procedure of the recognizer.

Visual speech is a natural candidate here, because it is independent of the acoustic environment. Also, evidence from human speech perception convincingly shows that visual cues might considerably contribute to speech comprehension. Not surprisingly, since the first attempt by Petajan in 1984 [3], a range of Audio Visual Speech Recognition (AVSR) systems has been developed, which confirmed the initial assumption that lip reading information is particularly helpful for recognizing noisy speech. Although there are clear differences in how these systems process audio and visual information and combine them together, they all share a quite similar system architecture based on a state-of-the-art approach to word recognition using phones as a subword modeling unit.

As an alternative, the use of articulatory features (AF) for ASR has been proposed [4], [5] and [6]. Articulatory features are usually described as abstract classes, which capture relevant characteristics of the speech signal in terms of articulatory information. These classes can be used as an intermediate representation, leading to a two-stage classification procedure with a remarkable degree of robustness under noisy conditions. Moreover, compared to purely acoustic features (like MFCC), AFs can also be used to represent properties of the speech production process, such as lip rounding, tongue position, manner of articulation, etc.

Although a number of ideas have been proposed for using AFs in conventional ASR systems, as far as we know, only few of them [7] and [8], have addressed the question of representing visual cues as AFs. Since there is an apparent correlation between some of the articulatory features and the visual shape of the lips during speaking, namely for labial consonants which are pronounced with closed lips and the roundedness feature which provides important cues to distinguish different kinds of vowels, articulatory features might lend themselves as an appropriate interface to integrate visual cues into the recognition procedure. This motivates us to investigate the possibility of combining these two lines of research in order to build a more robust speech recognition system.

We propose a two-stage architecture (see Figure 1), where abstract articulatory classes are extracted in parallel from both the speech signal and the video frames by means of statistical classifiers. The second stage then combines their outputs into AF-tuples and maps them to a corresponding phone stream. Finally, a lexical search maps this stream to words sequences as output. Such an architecture will not only facilitate the use of articulatory information within the speech recognition system, but also allows us to investigate to which degree the acoustic and visual contributions to phone classification are complementary. In this paper, we first provide a background review of previous work on AVSR in section 2. In section 3 we introduce the AF-based AVSR system and our N-best decision approach. Experimental results and conclusion are finally given in section 4 and 5 respectively.

2. Previous Work in AVSR

In AVSR, knowledge from diverse areas has to be brought together in a single decision procedure to successfully integrate the available evidence from the two input channels. In general,

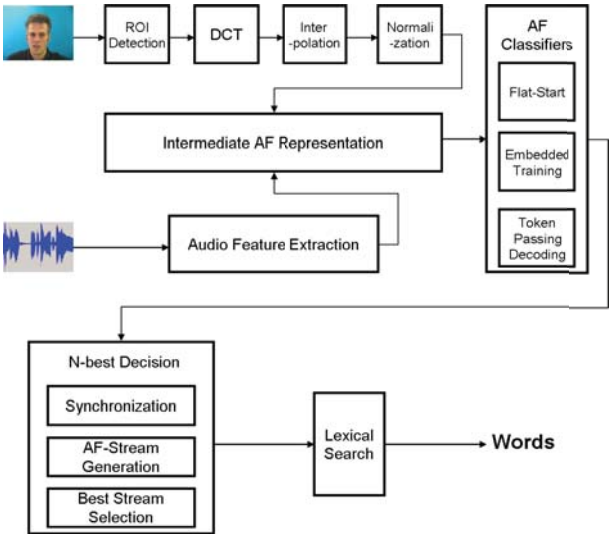

Figure 1: *AF-based two-stage AVSR architecture.*

five major design issues can be identified for such a system: 1) the design of the audio feature front-end, 2) the design of the visual feature front-end, 3) the choice of audio-visual speech classes, 4) the audio-visual speech classification, and 5) the fusion of audio-visual information. Many publications have addressed issues concerning the visual feature front-end, and the problem of audio-visual information fusion.

Opposed to the audio feature front-end, the visual one has to deal with extracting informative visual speech features in a robust manner. For that purpose, the raw video data of the speaker are firstly preprocessed to detect and extract the region of interest (ROI), namely the mouth region. Then, different algorithms can be employed for converting the ROI into feature vectors for further computing. To obtain useful visual features, three different approaches are available: the appearance-based, the shape-based and the hybrid method. Appearance-based features make use of all pixel level intensity and color values within the ROI. To reduce the extremely high dimensionality of these feature vectors various algorithms, like DCT, PCA, LDA, etc. can be used. Shape-based methods, on the other hand, are based on a model of the lip contour. The visual features are then chosen from the parameters of these models. Well known approaches in this category are models based on simple parameters as mouth height and width, snake models, and Active Shape Models (ASM) [9]. Finally, the hybrid method combines the two approaches by considering both pixel and shape evidence. In particular, approaches like the Active Appearance Model (AAM) have been shown to considerably improve the performance of shape-based features [10].

For integrating the different contributions from the two channels of an AVSR system two fundamentally different types of fusion techniques, namely feature-level and decision-level fusion, can be used. The feature-level fusion combines audio and visual features into a single feature vector [11]. Then, one classifier is trained based on those vectors. Decision-level fusion, on the other hand, trains two individual classifiers based on audio and visual feature data. The integration of their results is attempted in a subsequent step and can be achieved at different representational levels (sub-phonetic, phone, word, or utterance) [12]. In our paper, we employ the decision-level fusion idea to integrate the AFs from audio and visual speech signal.

3. AF-based AVSR framework

3.1. Articulatory Features

The basic idea of the AF approach is to use an additional speech signal representation situated between the acoustic signal pre-processing level and the subword unit probability estimation level. This representation is composed of articulatory features, i.e. abstract classes describing articulation-related information which is deemed relevant for the distinction between speech sounds.

Several reasons make AFs attractive for ASR. Firstly, they can provide a rather detailed description of coarticulation phenomena, since they are related to both, the acoustic signal and the higher level of linguistic information. In particular, they are able to accommodate the kind of asynchronous transitions between subsequent segments that can be observed with articulatory movements. Secondly, compared to a phone-based classification system, the parallel independent AF-based classification system makes use of fewer classes, which therefore are better suited to be used in case of sparse training data. Table 1 lists the articulatory features used for our experiments.

Table 1: *AFs used in AVSR framework.*

Features	Values	Num. Classes
Voicing	voiced, voiceless	2
Rounding	round, nil, flat	3
Manner	vowel, nasal, lateral, approximant, fricative	5
Place	dental, labial, retroflex, velar, high, mid, low	7
Front-Back	front, nil, back	3
Visual opening	open, close	2
Visual rounding	round, nil, flat	3

3.2. AF Recognition

Previous work for recognizing articulatory features from the speech signal has used Artificial Neural Networks, Hidden Markov Models, linear dynamic models and dynamic Bayesian networks. In our experiments we have concentrated on the use of Hidden Markov Models. In contrast to [4] we do not attempt to detect the articulatory features in a pure bottom-up fashion, but train a number of independent word recognizers, where the words are defined in terms of AFs instead, as usual, in terms of phones. These word recognizers are then applied in parallel to the audio or video data and their outcomes are word sequences which can also be interpreted in terms of sequences of articulatory features. This approach has the advantage that it allows us to integrate higher level information from a language model already during the AF-detection phase.

Seven AF-based models (5 from the audio signal and 2 from the video stream) have been trained by Baum-Welch reestimation. Instead of selecting the single best decoding

result, we determine the N-best hypotheses for all the AF-based classifiers. For recognition the Token Passing algorithm [13] is used. Token Passing saves the best tokens at each word boundary, which gives the potential for generating a lattice of hypotheses rather than only a single best hypothesis. Since the tokens are saved at the word level, the output is actually a sequence of loosely synchronized hidden words. They are emphasized as loose synchronization, since the HMM embedded training cannot guarantee a strict synchronization of the AFs within a word. However, thanks to the short pause models, which are usually easy to train, word boundaries can be rather reliably detected during recognition. In accordance with this observation, we are able to represent the recognized words as AF sequences, which are force aligned according to the word boundaries.

3.3. N-best Decision

The output of the AF-based recognizers will be processed in the second stage with the goal to combine the various channels into a single sequence of AF representations for which a meaningful phone representation exists. For this purpose, we propose an N-best decision scheme which computes the results of the first stage classifiers into a number of coherent AF tuples, which can be mapped to the phones contained in a code book. This approach is similar to the mixture of experts (ME) architecture proposed by [14]. Having available, however, the N-best output from the first stage, it seems more likely that the optimal results will be taken and a more reliable mapping between articulatory features and phone representations can be established. In our experiments, we have always chosen the five best hypotheses.

The N-best decision schema is invoked after the decoding stage of the AF-based classifiers. It consists of five procedures, namely 1) Synchronization, 2) AF tuple generation, 3) Best output selection, 4) Weighting and 5) Lexical Search. The input of the N-best decision schema is taken from the first stage N-best outputs, which are decoded word level sentences. Based on the idea of "loose synchronization", the best sentence decision problem is converted into a best word decision scenario. For that purpose word sequences are synchronized by normalizing the word length according to a majority vote among all the available output candidates of the AF-based classifiers. The length of a word defined here refers to the number of AF segments within that word. The normalization becomes necessary since the output of the first stage might be incoherent between alternative recognition hypotheses of the same classifier and across the different articulatory channels. It is carried out as a greedy search and assumes that (1) word boundaries can be reliably detected and (2) the word length is roughly comparable, two conditions which are fulfilled for the data in the GRID corpus. By selecting the optimal length, actually those word hypotheses are excluded, which are only supported by a minority of AF-based recognizers.

3.4. AF tuple Generation

AF classes can be combined into AF tuples where each component of the arity tuple of the tuple corresponds to the number of AF-based recognizers of the first stage. While some AF tuples can be mapped to phones, others can not. E.g. the tuple `[voiceless, fricative, labial, nil, nil]` can be mapped to the phone [f]. We maintain the possible mappings in a manually created AF-to-phone table.

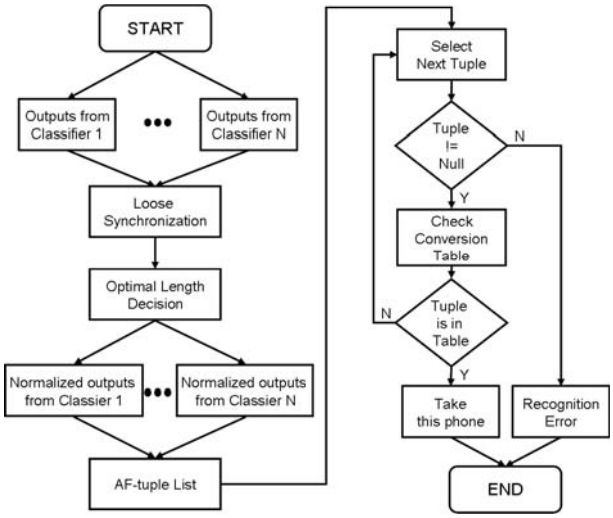

Figure 2: *The flow chart of N-Best Decision Schema*

AF tuples are generated according to the scores from the first classifiers. The first AF tuple, for example, is generated by combining the topmost candidate decision from all the classifiers. The second one will replace the topmost candidate of the most unreliable classifier by its second best choice, etc. Since we have only chosen the five best candidates of the AF-based word recognizers and in many cases they agree in the proposed recognition results, we are able to consider all combination possibilities when generating AF tuples.

Ideally, the AF tuple derived from the best decision of each classifier is the most likely one in all candidates. However, since the combined results are based on inaccurate first stage classifiers, it might not always be possible to map the parallel feature assignments into phones. Therefore, we need to exclude such combinations from consideration. Figure 2 shows the flow chart of N-best decision schema. The right part indicates the logic of best output selection. If the first output from the N-best list cannot be found in this table, it is replaced by another one from the list. If none of these tuples can be mapped into phones, a recognition error will be generated. The generated AF tuples are ranked based on the accumulated confidence score.

A phone stream is then defined as a sequence of phones which are admissible according to the AF-to-phone table. Eventually, this phone stream will be mapped into words according to phone-to-word table. For this purpose, a pronunciation dictionary including some pronunciation variants is used. For instance, "five" is transcribed both as [f ai v] and [f ai f].

Weighting is used in both, the synchronization and the AF tuple generation step. In order to vote for the optimal length of a word in the synchronization step, the decision could be weighted according to the recognition accuracy of the AF-based classifiers. A similar weighting scheme can be applied for generating the AF tuples. In our experiments the audio-based manner and place features have been assigned

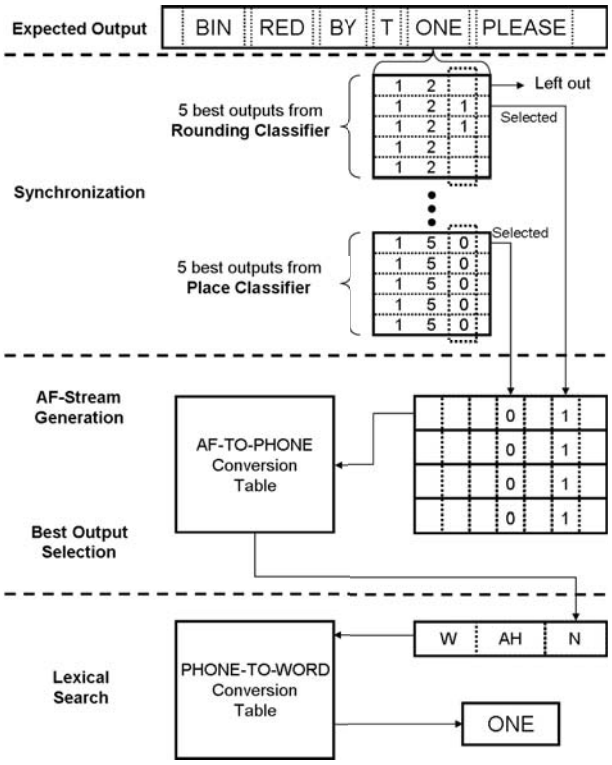

Figure 3: *Example for the N-best decision schema. AF-classes are coded as numbers*

a higher weight as compared to the other AF classification results.

The example shown in Figure 3 illustrates the data flow within the N-best decision schema. The testing data is the sentence "bin red by t one please". Since the word boundary is reliably detected by all first level AF-based word recognizers, the synchronization can be achieved word by word. When the word "one" is processed, the five best decisions of the rounding classifier have a different length of words. Voting determines the "optimal" length among all word candidates from all AF classifiers to three segments and all candidates with another length are no longer considered. After this synchronization step, we select the AFs from all classifiers and combine them into AF tuples. Their number is limited in our example, because all the AF-tuples have the same value in this example, which can be mapped to the phone [n]. Together with the two neighboring phones eventually the word "one" is decoded.

4. Experiments and Results

We performed an initial experiment on the GRID corpus [15], which is a continuous audio-visual speech corpus for an English small vocabulary task. It contains 1000 sentences spoken by each of 34 speakers. The original audio and video data were recorded under clean acoustic conditions, and the video shows only a frontal view of each subject's face. The sentences in GRID are speech commands according to a very simple grammar. The total of 51 words within the vocabulary consist of 4 command words, 4 words representing color, 4 prepositions, 26 letters, 10 digits and 4 adverbs.

In the audio channel, the raw speech signal was converted into a sequence of vector parameters with a fixed 25ms frame and a frame rate of 10ms. The 12 dimensional MFCCs were then obtained and one extra dimension was added as normalized log energy. Finally, the 13 dimensions parameters are expanded to 39 dimensions by adding first and second order derivations. On the visual processing side, the mouth region within a rectangular window was detected as ROI. This was done by applying a classifier trained by the rapid and robust Viola-Jones object detection algorithm [16]. The video was recorded as a sequence of images with a frame rate of 40ms. These colored images are further transformed into gray-scale ones. By using appearance (pixel) based method, every pixel inside the detected ROI images was considered as a feature.

To make classification feasible, we decrease the number of dimensions using the DCT transform. The final visual feature vectors contain 26 of the highest energy components. In order to compensate for the frame rate difference, the visual features were further interpolated from 25Hz to 100Hz. The final visual features can be trained either individually or combined with MFCC features.

In order to determine the articulatory information, we then classify the low level features into articulatory classes using left-right HMMs with 3 emitting states. The models are initialized with the flat start method [17] according to the features in each channel and the HMM parameters are trained with maximum likelihood estimation. After ten iterations, the models were finally expanded to 16 mixtures except for the manner tier where the models are expanded to only 8 mixtures. For the recognition, the token passing algorithm is used without any pruning factor. We take the five best outputs from each individual channel and apply the N-best decision scheme described in Figure 3 to them.

Figure 4 shows a comparison of three AF-based recognition systems with respect to their accuracy. Compared to the results of [5], where the articulatory features are also trained by HMMs, our system obtained better results in all individual classifiers. This performance is comparable to the one reported in [4], where Kirchhoff trained MLPs as AF classifiers. Since the first stage of our system actually are AF-based word recognizers, the figure also presents the corresponding word recognition accuracy in the different channels. It is considerably lower than AF accuracy because certain words can hardly be decoded using only AF classes.

For comparison to a conventional ASR, we trained a baseline phone-based classifier by means of HTK. This baseline ASR is using left-right HMMs with 3 emitting states single mixture models. The monophone HMM models were further extended to context dependent triphone models. The test data is then recognized with the Viterbi algorithm using a simple language model. Table 2 shows the word recognition accuracy results for different versions of the second stage. As expected, the word recognition accuracy of the individual first stage AF-classifiers is considerably lower than that of the baseline triphone-based recognizer. After applying the N-best decision schema to combine the AF-decoding results, however, the overall word recognition accuracy rises even above that of the phone-based approach.

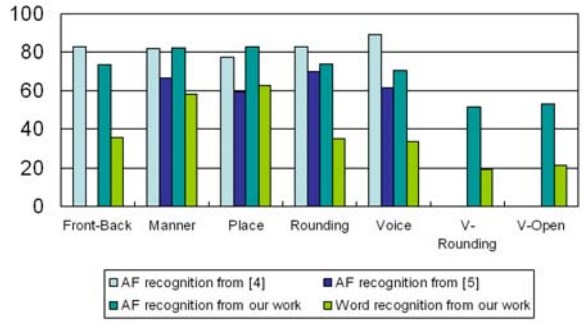

Figure 4: *The first stage feature accuracy rates comparison on three AF-based recognition systems*

Table 2: *AF-based word recognition accuracy after N-best decision*

Audio AFs only	Audio-Visual AFs	Phone-based
93.57	93.71	90.34

5. Conclusions

In this paper, we have proposed an audio visual speech recognition framework based on articulatory features. Within such a two-stage architecture, the multi-channel AF-classifiers are trained and tested with either audio or video data. Their results are combined in a second stage, using an N-best decision schema. Although the individual multi-channel AF-classifiers have a word-level performance far below the phone-based accuracy, their combination by the N-best decision schema is able to outperform the phone-based approach. When comparing the audio-only and audio-visual AF-based systems, the latter one gives slightly better results even under clean acoustic conditions and by further fine tuning the visual preprocessing algorithms a further improvement can be expected.

The experiment scarries out so far, confirm the potential of the approach for combining acoustic and visual cues for speech recognition purposes. So far, however, the algorithms for AF stream synchronization, best tuple selection and lexical search are rather simple ones. In addition to improving the visual components, more sophisticated solutions will be needed, if the approach is to be ported to more ambitions application domains. Possible candidates are statistical approaches like SVM or HMM, respectively.

6. Acknowledgements

This research was supported by the German Research Foundation (DFG) and the Ministry of Education of the Peoples Republic of China through the CINACS (Cross-Modal Interactions in Natural and Artificial Cognitive Systems) research school.

7. References

[1] K.-F. Lee, “Large-vocabulary speaker-independent continuous speech recognition: The Sphinx system”, Ph.D. Thesis, Carnegie Mellon University, 1988.

[2] R. Bakis, et al., “Transcription of broadcast news shows with the IBM large vocabulary speech recognition system”, proceedings of the Speech Recognition Workshop, 1997,67-72,1997

[3] E. D. Petajan, “Automatic lipreading to enhance speech recognition”, in Proc. Global Telecomm. Conf., Atlanta, GA, 1984, pp. 265C272.

[4] K. Kirchhoff, “Robust Speech Recognition Using Articulatory Information”, PhD thesis, University of Bielefeld, 1999.

[5] T. Abu-Amer and J. Carson-Berndsen , “HARTFEX: A Multi-Dimensional System of HMM Based Recognizers for Articulatory Feature Extraction”, In Proc. Eurospeech. Geneva, Switzerland. 2003.

[6] M. Wester, J. Frankel, and S. King, “Asynchronous articulatory feature recognition using dynamic Bayesian networks”, in IEICI Beyond HMM Workshop, 2004.

[7] K. Saenko, T. Darrell, and J. Glass , “Articulatory Features for Robust Visual Speech Recognition”, Proc. ICMI, State College, PA, October 2004.

[8] K. Livescu et al., Articulatory feature-based methods for acoustic and audio-visual speech recognition: JHU Summer Workshop Final Report, Technical report, Johns Hopkins University Center for Language and Speech Processing, 2007

[9] M. E. Hennecke, D. G. Stork, and K. V. Prasad, “Visionary speech: Looking ahead to practical speechreading systems”, in Speechreading by Humans and Machines, D. G. Stork and M. E. Hennecke, Eds. Berlin, Germany: Springer, 1996, pp. 331C349.

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor, “Active appearance models”, in Proc. Europ. Conf. Computer Vision, Freiburg, Germany, 1998, pp. 484C498.

[11] R. Goecke, G. Potamianos, and C. Neti, “Noisy audio feature enhancement using audio-visual speech data”, in Proc. Int. Conf. Acoust., Speech, Signal Processing, Orlando, FL, May 13C17, 2002, pp. 2025C2028.

[12] G. Potamianos et al., .Automatic recognition of audio-visual speech: Recent progress and challenges,. Proc. IEEE, vol. 91, no. 9, 2003.

[13] S. Young, N. Russell and J. Thornton, “Token Passing: a Conceptual Model for Connected Speech Recognition Systems”, CUED Technical Report F INFENG/TR38, Cambridge University, 1989.

[14] R.A. Jacobs, M.I. Jordan, S.J. Nowland, and G.E. Hinton. Adaptive mixtures of local experts. Neural Computation, 3:79C87, 1994.

[15] J. Barker, M. Cooke, S. Cunningham and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition”, Journal of the Acoustical Society of America, 120: 2421-2424,2006.

[16] P. Viola and M. Jones, “Robust Real-time Object Detection”, IEEE International Journal of Computer Vision vol.57, no.2, pp.137-154, May 2004.

[17] S. Young, G. Evermann, M. Gales, et al. Cambridge University, “HTK Book 3.4”, http://htk.eng.cam.ac.uk/, 2007.