Direct, Modular and Hybrid Audio to Visual Speech Conversion methods – a Comparative Study

Gyorgy Takacs

Faculty of Information Technology, Peter Pazmany University, Budapest, Hungary

takacs.gyorgy@itk.ppke.hu

Abstract

A systematic comparative study of audio to visual speech conversion methods is described in this paper. A direct conversion system is compared to conceptually different ASR based solutions. Hybrid versions of the different solutions will also be presented. The methods are tested using the same speech material, audio preprocessing and facial motion visualization units. Only the conversion blocks are changed. Subjective opinion score evaluation tests prove the naturalness of the direct conversion is the best.

Index Terms: facial animation, audio to visual conversion

1. Introduction

The audio to visual speech (ATVS) conversion targets to convert audio speech into visual speech. There are different concepts, some of them use automatic speech recognition (ASR) to extract phonetic information from the signal and on the phoneme string use some kind of visual coarticulation rule set or model in a range from simple viseme interpolation to phoneme viseme cross influencing sophisticated models[1].

One of the main properties of ASR based solution is the possibility of using language models. There are semi ASR based approaches also like [2] using phoneme level probabilities without language level. Other approaches among others use direct conversion between the modalities by a learning system [3, 4] without utilizing phoneme level information.

Research laboratories develop solutions, and the evaluations of the solutions are intelligibility tests and/or opinion score tests. The individual results are independent from each other so it is hard to tell which approach is better than the other. Now we describe a comparative evaluation which is performed between different conversion approaches by keeping all the other components of the workflow to be the same. See Fig 1.

We also introduce a hybrid method of different concepts which performs well as a result of our opinion tests. There are different aspects of quality of ATVS systems. The best possible conversion regarding intelligibility makes lip-reading possible. We traditionally work with hearing impaired, so intelligibility can be tested in this term. The best possible conversion concerning naturalness makes output which can not be distinguished from a record of original facial motion.

In our task we start from a natural acoustic speech signal, and for all speech qualities the best possible visual speech should be generated. In other words, translation between the modalities should be done, independently of the speech quality in terms of articulation, coding and noise ratio.

2. The compared systems

Basically five kinds of approaches have been evaluated:

- a reference natural facial motion
- a direct conversion system
- an ASR based cartoon industry ad-hoc standard method
- a modular system of ASR and text based sophisticated visual coarticulation modeling
- and the hybrid of modular ATVS and direct conversion.

The frontend of the methods up to the ATVS con-version is common. The same voice data is processed. Two speakers are in the database. One of the speakers is used for training, the other is for testing.

The visualization of the output of the ATVS methods is also common. The results are represented by facial animation parameters (FAP). The FAP is a part of MPEG-4 standard. Each FAP data flow is animated on the same head model. There are better facial descriptors than MPEG-4 but our motion capture system could not give more detail than MPEG-4, so we used this widely popular system, and simplified the more sophisticated methods to this common space. MPEG-4 FAP is a facial animation coding standard, it represents normalized facial feature point displacements.

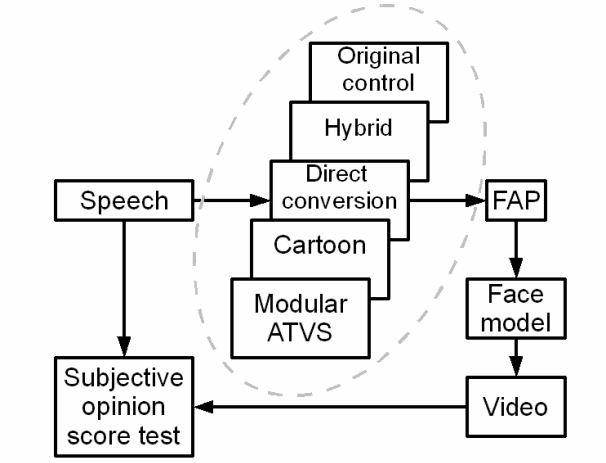

Figure 1: *Multiple conversion methods were tested in the same environment..*

The videos used in the tests are created from FAP sequence by an Avisynth [5] 3D face renderer plugin, which provides a convenient way of testing control parameters and handling multiple videos from multiple sources with cross-referencing trimming intervals, which is the usual task of subjective test compilation.

2.1. Original facial motion

Audiovisual speech was recorded. A video record contains the face of a subject with markers. The markers were tracked in 2D pixel space. Following the MPEG-4 standard the model's facial units (FAPU) were calculated in this pixel space, and the facial parameters were coded in FAP format using this FAPU. This method makes facial parameters head independent according to the MPEG-4.
Some automatic tracking errors were corrected manually and slight noise ¯ltering was done on the coordinates before FAP coding. The synchrony was checked by short-term audiovisual signal and bilabial plosives.
Original voice signal was used as input for the ATVS methods.

2.2. Direct conversion

Our research group developed a direct conversion system[3] published in 2006. The direct conversion takes actual voice segment and by a machine learning component estimates the best articulated facial parameters for the voice. It does not use phoneme or viseme level nor language-dependent elements. The machine learning method is usually regression by examples of audio and video data pairs. See Fig 2.
In our case a backpropagation neural network was trained between carefully chosen representations of the audio and the video data. MFCC was performed on the audio data for each frame of the database. The facial data was expressed with Principal Component Analysis (PCA) of the FAP. The neural network used 11 frame long window on the input side (5 frames to the past and 5 frames to the future), and 4 principal component weights on the output.
The trained neural network was used on the test set with a different speaker. Choosing training speaker is an important detail. Our speaker is a professional lip-speaker who works with deaf and hard of hearing people.

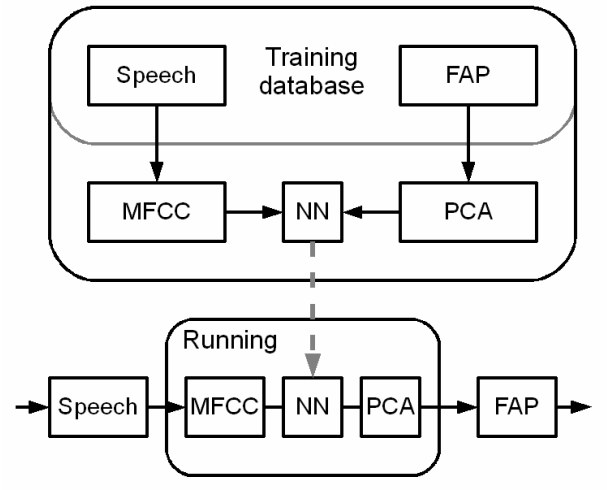

Figure 2: *Structure of direct conversion.*

2.3. ASR based solutions

For the ASR based approaches we used the best available speech recognition system for Hungarian. This is a piece of work of Mihajlik et al. The system is capable to use language model or vocabulary, and during the test we used both informed and uninformed recognitions. Uninformed recognition uses only general properties of the language, informed recognition uses vocabulary with the words occur in the test material.

In the ASR, a standard frame synchronous Weighted Finite State Transducer - Hidden Markov Model (WFST-HMM) decoder called as VOXerver [6, 7] was applied to obtain the phonemic segmentation of input waveforms. MFCC based feature vectors were computed with delta and delta-delta components. Blind channel equalization was used in the cepstral domain to reduce linear distortions as in [8]. Speaker independent cross-word decision-tree based triphone acoustic models were applied, trained previously on the MRBA Hungarian speech database [9].

In the uninformed ASR system, phoneme-bigram phonotactic model constrained the decoding process. The phoneme-bigram probabilities were estimated on the MRBA database. In the informed ASR system a zerogram word language model was used with a vocabulary size of 120. Pronunciation of words were determined automatically as described in [10].

In both type of speech recognition approaches the WFST-HMM recognition network was constructed o²ine using the AT&T FSM toolkit [11]. In the case of the informed system, phoneme labels were projected to the output of the transducer instead of word labels.

The system gives 10 ms precision of segmentation and the most probable phoneme for each segment. This data will be used to create visual speech.

2.3.1. Cartoon control

The baseline solution is the animation industry ad-hoc standard phoneme-to-phoneme interpolation with directly linked visemes. This approach is particularly popular among cartoon animators since the viseme count can be taken into consideration, the variety of the used visemes is scalable. The animator uses a table to connect phonemes to visemes. In the best case all the visemes for the given language is present. The interpolation of the visual states is uniform in this case. The sophisticated visual blending is supported by the next contestant.

In this solution the viseme set was the full value set of the ASR, the sample visemes were extracted from the original facial motion and linear interpolation was used.

2.3.2. Modular ATVS

We call an ATVS system modular ATVS (MATVS) if it consists of a separable ASR subsystem and a phoneme string to visual speech synthesizer subsystem. This is a particularly popular approach, since ASR technologies are well developed, standalone trainable and testable. See Fig 3.

As of 2009 the best available ASR is combined with the best available text based visual coarticulation sys- tem for Hungarian by Czap et al called TTVS [12].

This is a text to visual speech conversion system, a part of a text to audiovisual system without the voice synthesizer component. The system's workflow consists of a text preprocessor, a phoneme-to-viseme mapping with phoneme neighborhood dependent effect ratio, filtering and other post-processing steps. We hijacked the system in the text preprocessor step by injecting readily time-aligned data.

TTVS features dominant, uncertain and mixed dominance classes according to the level of influence by the neighborhood, and uses a database of mixed and uncertain class members' behavior in different neighborhoods.

The ASR system gives an output of phoneme strings with timing information. TTVS produces a high quality video from this data using Poser which was processed as an original recording, and FAPs were extracted. To test the methods we had to use the same virtual head, please note that some of the valuable information was lost during the conversion, so the test result may not show the real quality of the whole TTVS system, only the conversion part and only for those parameters which can be handled in our MPEG-4 subset.

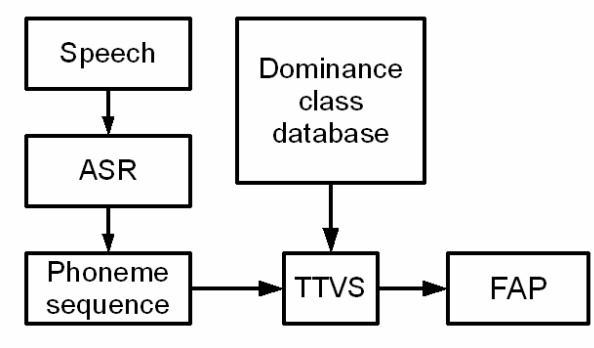

Figure 3: *Modular ATVS consists of an ASR subsystem and a text to visual speech subsystem.*

2.4. Hybrid control

Results of different systems are in a common linear space, since all of them are represented in FAP. Therefore the average of different controls is an appropriate control also. We used an inverse amplitude weighted mean of the output of the direct conversion and the output of uninformed modular ATVS. Weighting is to equalize the different articulation amplitudes in the result. The only synchronization between the control parameters is the common voice source.

3. Tests and results

We made subjective opinion score tests with the video material created by each of the methods and the original audiovisual recording. For modular ATVS tests we used both informed and uninformed recognition results, which are detailed below.

3.1. ASR subsystem

The quality of the recognition has two aspects. One of them is the precision of the assumed phoneme string. This is 100% at the informed run since the test set consists of small set of words as names of months or digits. The uninformed run falsely recognizes phonemes in 25.21% of the video frames. This may seems too high error ratio, but an ATVS using this input performes surprisingly well. The reason of this phenomenon may be the special confusion pattern which makes the error small if it is expressed with the resulting visual data. The speech recognizer confuses phonemes with visemes closer to each other more frequently than with others. The error expressed in relative viseme distance using quadratic metrics in FAP space is only 9.6% compared to random confusions for the whole data, or if we count only the falsely recognized frames it is still 52% of the random confusions. The other point of view is the precision of the segmentation. This was a bit harder task to the speech recognizer system. The uninformed run was more precise on the average than the informed. This makes a very heavy impact on the subjective opinion scores.

3.2. Opinion scores

The 58 test subjects evaluated the naturalness of the mouth motion. One of the original facial motion driven face models was shown as one of the bests and one of the most unaligned recognition based linearly interpolated motion as one of the worst. The test subjects were instructed to give scores between 1 and 5 according the presented videos.

The test material consisted of 7 videos of 5 methods. Another 5 videos synthesized from original facial motion recordings were added. The videos were presented in random order. Each video contained 2-4 separated words, started and ended in closed mouth state. The voice source of the direct conversion and the hybrid method was recorded with a person whose voice is not included in the training set of the neural network.

We decreased the articulatory amplitude of the direct conversion which was used for deaf people and has bigger opened mouth than the average.

The results (Table 1) confirm the theory of the beneficial properties of hybridization. The improvement of the subjective opinion score average between MATVS-2 and Hybrid-2 is significant with p = 0:00026 according to two-sample t-test. The advantage of direct conversion against MATVS is on the edge of signifycance with p = 0:0512 as well as the difference between the original speech and the direct conversion with p = 0:06 but MATVS is signifiantly worse than original speech with p =0:00029. The naturalness the excessive articulation is not eligible.

Table 1. *Results of opinion scores.*

Method	Average score
Original facial motion	3.73
Direct conversion	3.58
Hybrid	3.48
MATVS	3.43
Hybrid-2	2.97
Cartoon control	2.73
MATVS2	2.67

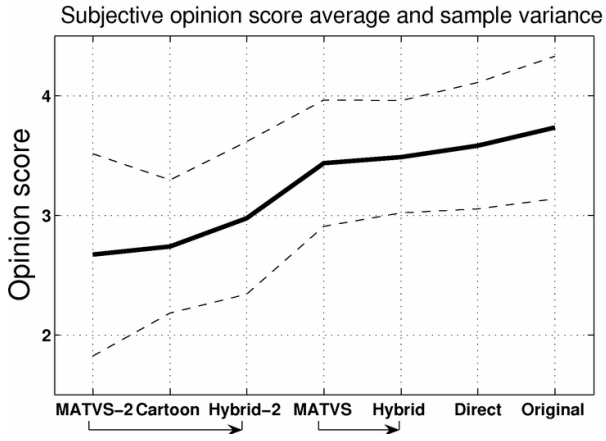

Figure 4: *Direct conversion is the closest to the original recording in the aspect of naturalness. Modular ATVS systems are vulnerable to synchronization errors, hybridization can help this*

4. Conclusions

A comparative study was proposed. Our direct conversion system was not compared to conceptually different conversion solutions before.

In the subjective tests we have the following definite results. We observed higher importance of the synchrony over phoneme precision in ASR based ATVS systems. There are publications on the high impact of correct timing in different aspects [12, 13, 14], but our result show explicitly that more accurate timing achieve much better subjective evaluation than more accurate phoneme sequence.

Also, we have shown that in the aspect of subjective evaluation, direct conversion is a method which produces the highest opinion score of 95.9% of an original facial motion recording with less computational complexity than ASR based solutions. We showed that hybridization can be used as a technique to significantly improve naturalness of segmentation problem oppressed ASR based ATVS systems.

For tasks where intelligibility is important (support for hearing impaired, visual information in noisy environment) modular ATVS is the best approach among the presented. Our mission of aiding hearing impaired people call upon us to consider using ASR based components. Hybridization worse intelligibility significantly, so only entertaining applications should use it. For naturalness (animation, entertaining applications) direct conversion and hybridization is a good choice.

5. Acknowledgements

The authors thank Péter Mihajlik for the support of the speech recognizer and the useful remarks, Láaszló Czap and Jáanos Mátyás for the TTVS, Márton Péri, PPKE ITK and its students who helped at the tests.

The research project was in the framework of Mobile Innovation Centre Budapest and supported by the National Office for Researc and Technology (Hungary)

6. References

[1] J. Beskow, I. Karlsson, J. Kewley, and G. Salvi. “Synface - a talking head telephone for the hearing-impaired”. Computers Helping People with Special Needs, 1178-1186, 2004.

[2] S. Al Moubayed, M. De Smet, and H. Van Hamme. “Lip synchronization: from phone lattice to pca eigen-projections using neural networks”. In Proceedings of Interspeech 2008, Brisbane, Australia, sep 2008.

[3] Gyorgy Takacs, Attila Tihanyi, Tamas Bardi, Gergely Feldhoffer, and Balint Srancsik. “Speech to facial animation conversion for deaf customers”. In 4th European Signal Processing Conf., Florence, Italy, 2006.

[4] Gregor Hofer, Junichi Yamagishi, and Hiroshi Shimodaira. “Speech-driven lip motion generation with a trajectory hmm.” In Proc. Interspeech 2008, pages 2314{2317, Brisbane, Australia, 2008.

[5] http://avisynth.org.

[6] Péter Mihajlik, Zoltán Tüske, Balázs Tarján, Bottyán Németh, and Tibor Fegyó.” Improved recognition of spontaneous hungarian speech morphological and acoustic modeling techniques for a less resourced task” IEEE Transaction on Audio Speech and Language Processing (submitted), 2008.

[7] P. Mihajlik, T.r Fegyó, B. Németh, and V. Trón. “Towards automatic transcription of large spoken archives in agglutinating languages: Hungarian asr for the malach project” In Speech and Dialogue: 10th International Conference, Pilsen, Czech Republic, 2007.

[8] P. Mihajlik, , Z. Tobler, Z. Tüske, and G. Gordos. “Evaluation and optimization of noise robust front-end technologies for the automatic recognition of hungarian telephone speech”. In Interspeech 2005 - Eurospeech: 9th European Conference on Speech Communication and Technology, Lisboa, Portugal, 2005.

[9] http://alpha.tmit.bme.hu/speech/hdbMRBA.php.

[10] P. Mihajlik, T. Révész, and P. Tatai. “Phonetic transcription in automatic speech recognition”. ACTA LINGUISTICA HUNGARICA, pages pp. 407{425, 2003.

[11] Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley. “Weighted ¯nite-state transducers in speech recognition” Computer Speech and Language, pages pp. 69{88, 2002.

[12] L. Czap and J. Mátyás. “Virtual speaker” Híradástechnika Selected Papers, Vol LX/6:pp.2{5, 2005.

[13] G¶erard Bailly, Oxana Govokhina, Gaspard Breton, and Fr¶ed¶eric Elisei. “A trainable trajectory formation model td-hmm parameterized for the lips” 2008 challenge. In Proceedings of Interspeech 2008, Brisbane, Australia, sep 2008. 7

[14] Gergely Feldhoffer, Tamas Bardi, Gyorgy Takacs, and Attila Tihanyi.” Temporal asymmetry in relations of acoustic and visual features of speech” In 15th European Signal Processing Conf., Poznan, Poland, 2007.