Adaptation of Talking Head System to a Different Language

Miloš Železný, Zdeněk Krňoul

University of West Bohemia, Faculty of Applied Sciences, Department of Cybernetics

Univerzitní 8, 306 14 Pilsen, Czech Republic, Email: zelezny@kky.zcu.cz,zdkrnoul@kky.zcu.cz

Abstract

This paper describes new techniques in order to adapt a talking head synthesis system for a different language. The synthesis system is originally designed for Czech language. Recently, experiments were performed on adaptation of the synthesis system to the English and Dutch languages. This paper generalizes the use of these procedures for other languages.

Introduction

Conversion of text to speech is increasingly used the electronic appliances and information systems. The conversion of text to visual speech (movements of mouth during speech) is a modality of text to speech system (TTS) specially designed for visual perception of speech. Visual speech is important at certain situations, because it is a conditioning factor for lip-reading. Humans often use lip-reading, especially when acoustic noise occurs. Lip-reading is also one communication means of deaf people. Artificially generated visual speech is expressed by computer animation of a human face. The animation is rendered in real time, or via video files. The acoustic as well as the visual component of speech differs for different languages [1]. The adaption of visual speech synthesis system to a new language is not always completely resolved.

Adaptation Method

The adaptation of talking head system can be summarized in the following steps: *acquisition of speech data*, *speech segmentation*, *data analysis*, *adaptation of the control model* and *adaptation of the animation model*.

Audiovisual Speech Database

The first step of the adaptation is capturing continuous speech of the target language. Firstly, we need to collect appropriate text material. We suggest to use phonetically balanced sentences. The length of sentences can vary from 3 words to 15 words and the number of sentences should be at least 500. Small number of sentences does not ensure proper coverage of co-articulation effect. For adaptation purpose, the speech data of one speaker is sufficient. The speaker, however, should have good articulation especially if we want to use final animation for lip-reading. Appropriate is to use a recording studio with low level of acoustical and visual noise. Unbalanced intensity of lighting and acoustic noise in the background cause corrupted audiovisual speech data.

The recording equipment consists of a video camera and microphone. The video camera has to be located approximately 5 meters from the speaker and lens is zoomed on front view of speaker's face. Optionally we can use a second camera or a mirror to capture in parallel the profile of the face. One channel and sampling frequency 44 kHz is sufficient for the acoustic component of speech. For the visual component, the resolution of 576x720 pixels (portrait) is optimal but the sampling frequency must be at least 25 frames per second (fps), or better 50 fps. The sources of light should have sufficient intensity due to the time of exposure. Lighting should not cause any shadow in the speaker's face and reflection, and suppress natural face colors. High contrast lips-skin and lips-teeth is crucial here. The cameras, or speaker's head should not move during the record. Important point is the synchronization of audio and video signals using the clapperboard, or an electronic synchronization device. The record during one day ensures to get same speaker's voice dispositions for all sentences.

Speech Segmentation

In general, speech segmentation divides the speech data into short intervals of phones. The segmentation process is implemented in two stages. In the first stage, time-synchronized audio-visual data are manually divided into sentences and pronunciation of words have to be corrected. The second stage automatically processes the sentences into speech segment corresponding phonemes, or visemes. For this purpose, we have to use a external segmentation tools, for example Hidden Markov Model Toolkit (HTK). We recommend to use the acoustic component only and all the phonetic units of the target language. The start boundary of each speech segment in the acoustic component determines then articulatory target in the synchronized visual component. Finally the recommended post-processing is manual inspection the speech segments. The sentences with error segmentation are discarded from the speech database, or boundaries of these segments have to be manually corrected.

Analysis of Image Data

Method of analysis of image data converts the mouth shape captured in video frames to a numerical representation (visual parameterization). We recommend using the method of Repeated template matching [2]. The analysis of image data is divided into three stages: *selection of templates*, *model of the mouth* and *processing of speech data*.

The first stage determines sample images of mouth shapes (the templates). The templates are manually collected from the video frames of the speech database. We

create the template by cut-out of mouth area in the video frame. We recommend to align the center of upper side of templates to the position of nose. The optimum number of templates is about 100, we use 101 templates for English, 83 for Czech, 117 for Dutch. Selected templates have to affect all major forms of mouth shape: different degrees of mouth opening, protrusion and width, contacts and interrelation of lips, teeth and tongue. However a larger number of templates cause very slow data analysis.

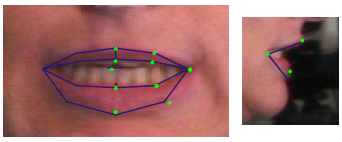

Figure 1: On left: one template and approximation of lip shape, on right: data from face profile (optional)

The second stage creates a model approximating mouth shapes in the templates. Firstly we manually get positions of control points in the templates. An approximation of lip contours is shown in Figure 1. The shifts of control points against the mean lip shape (closed lip) collected from all templates are processed by Principal Component Analysis (PCA). The principal components depend on collection of the templates but often model these articulatory parameters: lip opening, raising and rounding (protrusion). The last stage automatically goes through all the video frames of sentences and determines the hight correlated templates. We know the values of the articulatory parameters for these templates. The articulatory trajectories are then created by cubic spline interpolation and they are used for adaptation of the control model. The lip model can be optionally extended about shift of lips in 3rd dimension or tongue positions. For this purpose, we can use the time-synchronized video frames capturing the face profile.

Control Model

Analysis of image data and segmentation of speech provides training data to adapt the control model. Segmented articulatory trajectories allow us to collect articulatory targets of all the phonemes from the speech database. The control model uses a technique of Selection of Articulatory Targets (SAT) [3]. This technique is based on training of the regression trees. One regression tree models one articulatory parameter of particular phoneme. The root of the regression tree contains cluster of all the relevant articulatory targets. Regression questions split this cluster to sub-clusters by Euclidean distance. In the principle, the regression questions have to include the immediate phonetic context (tri-phone context). Optionally we can extend the set of questions about wider phonetic context, durations of the phoneme, etc. For large speech database (higher number of training sentences), regression trees can be pruned to achieve optimal size and robustness.

Animation Model

Animation model converts articulatory trajectory to visual speech. We assume 3D animation model based on pseudo-muscle facial deformation [4]. An illustration of animation model is shown in Figure 2. Adaptation of animation model consists of re-modeling new key mouth shapes similar to mouth shapes obtained during the analysis of image data. The animation model compute new mouth shape as their weighted combination. The correctness of adaptation can be validated on the templates. The animation model must produce the same mouth shapes.

Figure 2: The animation model originally developed for Czech language

Summary and Conclusions

The adaptation of visual speech synthesis system was tested on three languages: Czech, English and Dutch. We have used the adaptation also for the sign language as well. In this case, the visual speech synthesis is used in the automatic synthesis of Czech sign language. Audio-visual perception tests confirm intelligibility of synthesized visual speech. Proposed steps of adaptation involve several manual stages. Future work may focus on extending of the current adaptation process in order to make it more automatic.

Acknowledgments

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/09/P609, and the Ministry of Education of the Czech Republic, project No. ME08106.

References

[1] Věra Strnadová. *Hádej, co říkám aneb Odezírání je nejisté umění*. GONG, Praha, 1998.

[2] Zdeněk Krňoul and Miloš Železný. Innovations in czech audio-visual speech synthesis for precise articulation. In *Proceedings of AVSP 2007*, Hilvarenbeek, Netherlands, 2007.

[3] Zdeněk Krňoul and Miloš Železný. A development of czech talking head. In *Proceedings of Interspeech 2008*, Brisbane, Australia, 2008.

[4] Zdeněk Krňoul and Miloš Železný. Realistic face animation for a Czech Talking Head. In *Proceedings of TEXT, SPEECH and DIALOGUE, TSD 2004*, Brno, Czech republic, 2004.