INFLUENCE OF PHONE-VISEME TEMPORAL CORRELATIONS ON AUDIOVISUAL STT AND TTS PERFORMANCE

Alexey Karpov[a], *Andrey Ronzhin*[a], *Irina Kipyatkova*[a] & *Miloš Železný*[b]

[a]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia;
[b]University of West Bohemia in Pilsen, Czech Republic
karpov@iias.spb.su; ronzhin@iias.spb.su
kipyatkova@iias.spb.su; zelezny@kky.zcu.cz

ABSTRACT

In this paper, we present a research of temporal correlations of audiovisual units in continuous Russian speech. The corpus-based study identifies natural time asynchronies between flows of audible and visible speech modalities partially caused by inertance of the articulation organs. Original methods for speech asynchrony modeling have been proposed and studied using bimodal ASR and TTS systems. The experimental results have shown that use of asynchronous frameworks for combined audible and visible speech processing results in improvement of the accuracy of audiovisual speech recognition as well as the naturalness and the intelligibility of speech synthesis.

Keywords: audiovisual speech, recognition, synthesis, multimodal processing, viseme

1. INTRODUCTION

Auditory and visual cues of speech naturally supplement each other and their combined processing helps to improve quality and performance both of automatic speech-to-text (STT) and text-to-speech (TTS) systems, resulting in synergy in many cases like the famous McGurk effect. However, audible speech units (phones as representations of phonemes in speech) and visible ones ("visemes" as introduced by C. Fisher, there is no any widespread term that distinguishes abstract visual units from realizations of them in speech) do not have full temporal synchronization in natural speech. It is mainly caused by the human's speech production system, inertance of the vocal tract and its articulatory organs results in the coarticulation phenomenon [1], which reveals itself differently on two speech modalities and causes some asynchrony between them.

Time lags between flows of phones and visemes in speech are language-dependant. For instance, these speech modalities are almost simultaneous in Japanese [10], but there exists a considerable asynchrony in English, especially in American English, which is characterized by the hyper-articulation of many speakers.

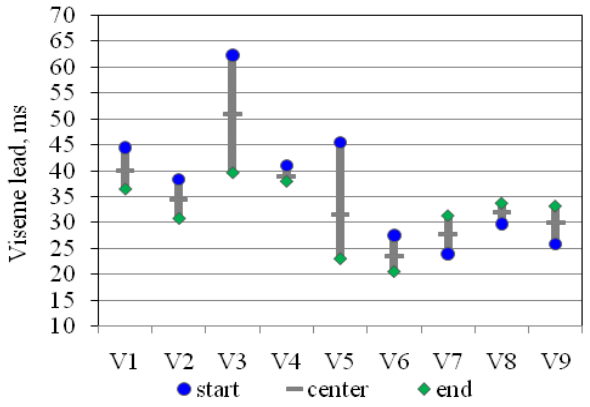
The present paper studies correlations between audible and visible speech units in Russian, which belongs to the Slavic languages. Since no appropriate bimodal Russian speech corpus was available, the audiovisual (AV) continuous speech database has been privately recorded and prepared. It contains pronunciations of phonetically-balanced sentences uttered by 10 native Russian speakers with normal articulation, both men and women 31 years old in average. The speakers uttered with normal speech tempo 1500 phrases in total consisting of 4-8 words each, including continuous sentences and connected digits. Sony DCR-PC1000 digital camcorder was used to capture video data with 720x576x25 fps and a built-in microphone located at 15-20 cm from the speaker's mouth was used for synchronous speech sound acquisition with 22 kHz sample rate, SNR ≈ 25 dB.

The bimodal database has been segmented semi-automatically by an automatic speech recognizer (ASR, Section 2) and checked by the expert way. Acoustic labeling contains 42 diverse context-independent phonemes of Russian speech corresponding to the SAMPA International phonetic alphabet. Labeling into phonemes and visemes was made one-to-one, i.e. each phoneme in the flow was associated with one viseme in order to keep the correspondence between the segmentations. Totally 10 diverse viseme classes (including pause V0 as a neutral lips configuration) are differentiated in Russian speech and they were used in segmentation of the visual data: V1 wide-opened mouth unrounded vowels – phonemes /a/, /e/; V2 the rest unrounded vowels – /i/, /1/ (in SAMPA notation); V3 rounded vowels – /o/, /u/; V4 bilabial consonants – /b/, /p/, /m/ (both hard and soft); V5 labiodentals – /f/, /v/; V6 alveolar

sonorants – /l/, /r/; V7 alveolar fricatives – /S/, /Z/, /tS'/, /S':/; V8 velar consonants – /g/, /k/, /x/, /j/; V9 dental and the rest cons. – /d/, /t/, /n/, /s/, /z/, /ts/.

After an analysis of both segmentations of the corpus, the temporal correlations between corresponding phonemes and visemes have been calculated and summarized in Figure 1, which shows mean values of visemes lead for start boundary, center and ending of the phone-viseme classes. If the phone starting delay is greater than the delay of the unit ending, then the corresponding viseme is more time-overlapping with the preceding phone (current viseme is extended); and on the contrary, if the phoneme starting delay is less than the ending delay, then the viseme corresponding to the following phone in the speech flow is more time-overlapping with the given phoneme (current viseme is reduced). As the result of the corpus-based study, the following issues of speech modalities asynchrony can be formulated: (1) visemes always lead in phone-viseme pairs; (2) at the beginning part of a phrase visual speech units usually lead more noticeably over the corresponding phonemes than in the central or ending part of the phrase; (3) greatest time lags are observed for the rounded vowel phones (up to 80 ms), a bit shorter for the bilabial obstruent consonants, and less for the remaining vowels; (4) the stressed rounded vowels have longer delays (sometimes over 100 ms for /u/) than the same unstressed vowel phones; (5) the best temporal correlations are observed for fricatives and sonorants, excluding bilabial /m/.

Figure 1: Mean values of viseme lead in AV units for start, center and end boundaries.



These issues coincide well with other studies, e.g. [3] reports that temporal asymmetry in relations of acoustic and visual features of spontaneous speech in the beginning of speaking can vary up to 100 ms and even more at slow speech tempo of ordinary speakers or professional lip-speakers.

2. AUDIOVISUAL SPEECH-TO-TEXT

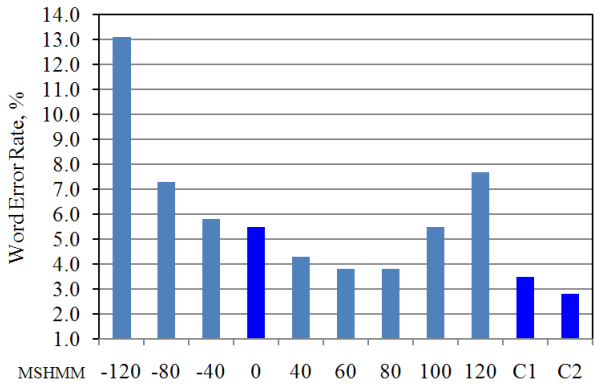
Some recent AV STT systems propose an asynchronous framework for speech decoding, for instance, using Coupled Hidden Markov Models (CHMM) [9] or Articulatory Feature Model [5]. These models are able to take into account some time lags in bimodal speech, at least inside the boundaries of AV units. However, there are no studies concerning influence of audio-to-video signal shifts on performance (accuracy and robustness) of speech recognition by state synchronous models. According to our hypothesis, the word error rate (WER) of AV STT should change with phasing signals relative to each other.

For the system training, we have used 60% speaker's utterances containing phonetically-rich phrases (90 sentences per speaker). The rest of the AV data consisting of utterances of 4-7 connected digits were used for the evaluation. As acoustic features, we used 12-dimentional Mel-Frequency Cepstral Coefficients calculated from 26 channel filter bank analysis of 20 ms long frames with 10 ms overlap, thus, the frequency of audio feature vectors is 100 Hz. The visual (articulatory) features are pixel-based with the Principal Component Analysis of the region of interest (lips and mouth area) in video frames upsampled from 25 to 100 Hz in order to correspond with the audio vectors frequency. Speech decoder was realized with HTK 3.4 Toolkit based on Multi-Stream HMMs (MSHMM) corresponding to all 42 phoneme/viseme units.

Figure 2 presents the WER in clean speech conditions (25 dB) for the MSHMM-based AV STT system, where first the audio data (feature vectors) and then the video data were delayed relatively to the other modality stream by 120/80/60/40 ms, respectively. The best results with the MSHMM-based recognizer have been achieved, when the stationary delay of the visual features stream was 40-80 ms (the WER is better by 1.7% than without (0) signal shift); on the contrary, a delay of the auditory features results in essential WER increase. These experiments have demonstrated the problem of asynchrony between auditory and visual speech features/units, and a short shift of the video data was able to increase the recognition rate of the state synchronous speech recognizer from 94.5 to 96.2%. These

results were compared with the CHMM-based approach (C1 in Figure 2). In all the previous experiments, WERs were worse than for the CHMM-based recognizer (3.3%). Optimal audio and video stream weights in these models were 1.4 and 0.6, correspondingly. This model was also compared to the model with individual viseme-dependent stream weights for each AV unit [6], which has demonstrated the WER = 2.7% (C2 in Figure 2). In the experiments with the CHMM-based recognizers, the signal phasing did not improve the WER that can be explained such model allows asynchronous speech decoding.

Figure 2: WER at various video delays.



3. TEXT-TO-AUDIOVISUAL SPEECH

There is a lack of research on modeling of natural phone-viseme temporal relations for AV TTS as well. Natural coherence of both speech modalities can be provided by 2D AV speech synthesis based on the multimodal unit selection approach [8]. Nevertheless, 3D model-based synthesizers, including concatenation-based and HMM-based systems, are usually not supplied with adequate asynchrony models. Among speech asynchrony models, embedded into bimodal TTS systems, we can point out the context-dependent phasing model [4]. In this phasing model, an average delay is associated with each context-dependent HMM. However, these delays do not take into account variability of synthesized speech rate.

We have implemented own 3D realistic talking head model for Czech and Russian, which is an audio-driven model, where the visual processing part is controlled by the results of audio TTS with the help of a modality asynchrony model. The talking head is based on a parametrically controllable 3D model of a head. Models of the visemes in the form of the sets of control points are concatenated to produce continuous stream of visual parameters. In our model, the coarticulation is modeled by the visual unit selection method in order to better achieve articulatory targets important for visual perception of certain phonemes (for example, occlusions in speech). The audio TTS is based on concatenation of allophones. Synchronization of face/lip movements with synthesized acoustical signal is based on timestamps of allophones in the synthesized speech flow. Duration of every allophone is based on allophone's average length and desired speech tempo. To model AV speech asynchrony and take into account different speech rates more, we have elaborated over 20 context-dependent timing rules for transitions between visemes [7].

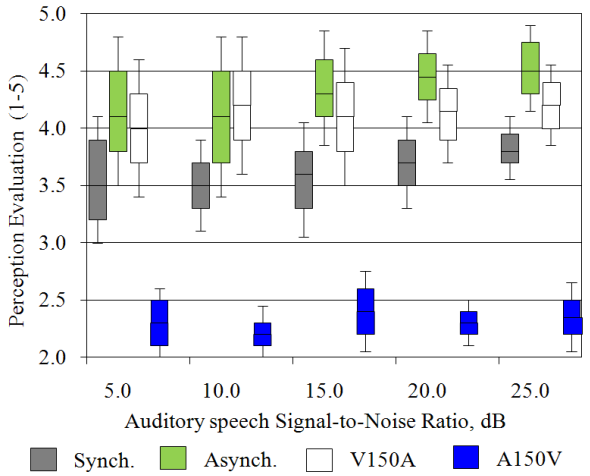
Some speech perception experiments with the talking head were made for evaluation of different kinds of modality asynchrony models implemented in the talking head by the criteria of speech naturalness and intelligibility. Three types of stimuli were applied: (1) auditory synthesized speech; (2) AV synthesized speech by the talking head with diverse synchronization models; (3) pre-recorded real speech/face (the same speaker was used for creation of the synthesized voice). Totally 20 phonetically-balanced continuous sentences were selected from the bimodal speech corpus and presented in a random order to informants. Each continuous phrase was composed of 5-7 well-known meaningful Russian words. However, all the selected phrases were meaningless on the whole or have a partial meaning so that to test human's visual and hearing perception without a-priory semantic knowledge.

10 volunteers of 20-35 years old with normal hearing and eyesight had to examine four kinds of AV speech synchronization models: (1) completely synchronous (phones and visemes in speech share same boundaries); (2) the talking head with the proposed asynchrony model using the set of timing rules; (3) a simple asynchrony model, where a stationary delay of 150 ms was applied to the synthesized audio signal relatively to the corresponding video signal (V150A); (4) a similar asynchrony model where a stationary delay of 150 ms was applied to the video signal (A150V). Moreover, babble ("cocktail party") noise with various intensity (SNR varied 25-5 dB) was added to the clean speech signal. The informants were asked to test the talking head and to evaluate the quality and naturalness of AV synchronization of the synthesized speech by the 5-point scale ("5" score is the best), comparing it with the real AV

speech recording as a reference sample. Also the informants had to write down a word string they recognized for each type of stimuli.

Box plots in Figure 3 show perception evaluations of AV speech synchronization models averaged over all the viewers, the results are statistically significant according to the analysis of variance. Most of the subjects have confirmed they see distinctions in the talking heads with different asynchrony models. Moreover, many respondents evaluated the completely synchronous model (Synch) with rather low marks; the majority of testers preferred the proposed original asynchrony model and one person of 10 preferred the V150A model. The informants were much more tolerant to video signal leading than to audio signal leading.

Figure 3: User perception evaluations vs. SNR.



Informant's evaluations decrease with decreasing SNR. An important measure is a distance between evaluation marks for all the models: it also decreases when SNR drops. So informants perceived differences between the synchronization models better in relatively clean speech, but in very noisy speech (SNR $\leq$ 10 dB) many informants did not catch any difference. However, the advantage in naturalness of the proposed asynchrony model in noiseless conditions was appreciated by the most of the informants.

It is known that influence of visual cue on speech intelligibility depends on language and environment [2]. In our experiments, the speech intelligibility of synthesized AV speech was lower than for real speech recordings in clean speech conditions. The intelligibility of the talking head was 87% and audio-only speech – 85%, while real speech provided the intelligibility over 98%. In environment with additive babble noise (SNR = 5dB), these models have demonstrated 53%, 38% and 84%, correspondingly. Statistically significant differences in the speech intelligibility of the talking head with different asynchrony models were not observed in the experiments.

4. CONCLUSION

We have proposed several methods for asynchrony modeling of AV speech flows in STT and TTS systems. A hypothesis, that the WER of bimodal ASR is changed with shifting AV speech signals relative to each other, has been confirmed and an optimal delay of video data in AV recordings is 40-80 ms in STT. Also the results of the speech perception with the 3D talking head prove that the proposed timing rule-based asynchrony model improves naturalness of synthesized AV speech.

This research is supported by the Grant of the President of Russia № MK-64898.2010.8, by the Ministry of Education and Science of Russia, project № 2579, and by the Ministry of Education of the Czech Republic, project № ME08106.

5. REFERENCES

[1] Browman, C.P., Goldstein, L. 1992. Articulatory phonology: An overview. *Phonetica* 49(3-4), 155-180.

[2] Chen, Y., Hazan, V. 2007. Language effects on the degree of visual influence in audiovisual speech perception. *Proc. 16th ICPhS* Saarbrücken, 2177-2180.

[3] Feldhoffer, G., Bardi, T., Takacs, G., Tihanyi, A. 2007. Temporal asymmetry in relations of acoustic and visual features of speech. *Proc 15th European Signal Processing Conference EUSIPCO* Poznan, 2341-2345.

[4] Govokhina, O., Bailly, G., Breton, G. 2007. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. *Proc. ISCA Speech Synthesis Workshop* Bonn.

[5] Hasegawa-Johnson, M., Livescu, K., Lal, P., Saenko K. 2007. Audiovisual speech recognition with articulator positions as hidden variables. *Proc. 16th ICPhS* Saarbrücken, 297-302.

[6] Karpov, A., Ronzhin, A., Markov, K., Zelezny, M. 2010. Viseme-dependent weight optimization for CHMM-based audio-visual speech recognition. *Proc. 10th Interspeech* Makuhari, 2678-2681.

[7] Karpov, A., Tsirulnik, L., Krnoul, Z., Ronzhin, A., Lobanov, B., Zelezny, M. 2009. Audio-visual speech asynchrony modeling in a talking head. *Proc. 9th Interspeech* Brighton, 2911-2914.

[8] Mattheyses, W., Latacz, L., Verhelst, W. 2009. On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing 2009*.

[9] Nefian, A.V., Liang, L.H., Pi, X., Xiaoxiang, X., Mao, C., Murphy, K. 2002. A Coupled HMM for audio-visual speech recognition. *Proc. ICASSP* Orlando, 2013-2016.

[10] Sekiyama, K. Tohkura, Y. 1993. Inter-language differences in the influence of visual cues in speech perception. *J. Phonetics* 21, 427-444.