INTERSPEECH 2009 BRIGHTON

Audio-Visual Speech Asynchrony Modeling in a Talking Head

Alexey Karpov[1], Liliya Tsirulnik[2], Zdeněk Krňoul[3], Andrey Ronzhin[1], Boris Lobanov[2], Miloš Železný[3]

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Russia
[2] United Institute of Informatics Problems of the National Academy of Sciences, Minsk, Belarus
[3] University of West Bohemia in Pilsen, Czech Republic
karpov@iias.spb.su, l.tsirulnik@newman.bas-net.by, zelezny@kky.zcu.cz

Abstract

An audio-visual speech synthesis system with modeling of asynchrony between auditory and visual speech modalities is proposed in the paper. Corpus-based study of real recordings gave us the required data for understanding the problem of modalities asynchrony that is partially caused by the co-articulation phenomena. A set of context-dependent timing rules and recommendations was elaborated in order to make a synchronization of auditory and visual speech cues of the animated talking head similar to a natural humanlike way. The cognitive evaluation of the model-based talking head for Russian with implementation of the original asynchrony model has shown high intelligibility and naturalness of audio-visual synthesized speech.

Index Terms: audio-visual speech processing, text-to-speech synthesis, multimodal speech perception, cognitive study

1. Introduction

A proper coherence between auditory and visual cues of speech is one of the most important problems in the field of audio-visual speech synthesis (AVSS). The essence of the question is that phoneme and viseme are not completely synchronized in the course of a natural speech flow. It is partially caused by co-articulation phenomena in speech production, i.e. influence of some preceding speech units on following ones and vice versa which intervenes both on visual and auditory speech cues. Co-articulation reveals itself differently on two speech modalities and causes asynchrony between them. This aspect has key significance for speech perception because it has an immediate influence on intelligibility and naturalness of real and synthesized speech. The degree of time asynchrony between flows of phonemes and visemes is different for different languages. It was found out, that they are almost simultaneous for Japanese [1], but considerable time lag between the two speech modalities is attested for English (especially for American English), which is characterized by rich articulation.

Essential attention is paid to bimodal asynchrony modeling for state-of-the-art audio-visual speech recognition systems. Several models were developed to handle this problem, such as Coupled Hidden Markov Models [2] or decision fusion models. However, there is a lack of research of asynchrony modeling for audio-visual text-to-speech (TTS) systems. Natural coherence of both speech modalities is well provided by 2D audio-visual TTS based on the multimodal unit selection approach [3]. Nevertheless 3D model-based synthesizers, including concatenation-based and HMM-based systems, are not usually supplied with any adequate asynchrony models.

Only recently researchers got down to investigation of the above-mentioned problem applied to AVSS. Among asynchrony models, embedded into real bimodal TTS-synthesis systems, we can point out a context-dependent phasing model that was proposed for French [4]. In the 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. During further research the phasing model was extended to Phased Hidden Markov Model [5].

In the theoretical model, presented in this paper, results of the above-mentioned investigations are taken into consideration for studying natural time discrepancy between auditory and visual speech units for Russian. Besides, an original synchronization model was proposed in order to improve both naturalness and intelligibility of synthesized Russian speech.

2. Corpus-based study of bimodal asynchrony

An audio-visual corpus of Russian speech was collected; it represents a phonetically-balanced text pronounced by 4 speakers, both men and women. All the persons are native Russian speakers with normal articulation at the age from 20 till 70 years. The phonetic content of the text was elaborated in such a way, that statistical coverage of context-dependent phonemes of the Russian speech and language was maximal. The recording session for each speaker lasted about 10 minutes. Sony digital video-camera in 720x576x25 fps mode was used to capture video signal and a high-quality stationary microphone, located at approximately 20 cm from the speaker's mouth, was applied for sound recording with the sampling rate of 22 kHz, mono, SNR > 35 dB.

The collected multimodal data were divided into auditory and visual parts by the modalities fission and single modalities were automatically segmented in terms of context-independent phonemes and visemes by a HMM-based audio-visual speech recognizer [6] using a well-known Viterbi-based forced alignment algorithm. However any automatic recognition-based technology is error-prone and the segmented data contained rather many mistakes in timestamps for labels, so the automatically-segmented speech corpus was manually examined and all the errors were corrected.

The bimodal database was segmented using labels of 42 context-independent phonemes of Russian speech corresponding to SAMPA International alphabet taking into account stressed variants of vowels and pause (silence). 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 correspondence between segmentations. As a result of the corpus-based study the following tendencies in asynchrony of modalities were discovered:

- Visemes always leading phoneme-viseme pairs, i.e. transition between two consequent visemes is made within the first phoneme. Although a few exceptions from this rule were found (in the case of rounded vowels before a pause), such exceptions can be discarded.
- The greatest time lag is observed for rounded vowel phonemes: /o/ and /u/ (above 80 ms), somewhat shorter one – for labial obstruent consonants: /v/, /f/, /p/, /b/ (40-60 ms), then for remaining vowels: /a/, /i/, /e/, /y/ (35-55 ms). For other phonemes the delays are considerably shorter.

- Stressed rounded vowel phonemes (/u_0/, /o_0/) have longer delays than the same unstressed phonemes. For the other vowels this is not valid.
- The best time coherence is observed for viseme-phoneme pairs of sonorants (/r/, /l/, /n/), excluding phoneme /m/.
- At the beginning of a phrase (1-st, 2-nd, 3-rd visemes) the visual units usually lead more noticeably over the corresponding phonemes than in the central part or at the end of a phrase.

Based on study of revealed asynchrony a set of context-dependent timing rules was defined (Section 3.3), which allows modeling of these phenomena for improvement of AVSS.

3. Talking head model for Russian

General architecture of the developed audio-visual speech synthesizer for Russian is depicted in Figure 1. In the given research we propose a way for implementation of an additional module in the standard architecture of audio-visual speech synthesizer, this is the module for modeling of asynchrony phenomena between corresponding auditory and visual units. The proposed talking head is an audio-driven model so the visual processing part is controlled by the results of text-to-speech system with the help of modalities asynchrony model. High-quality auditory speech synthesizer was designed for Russian especially, while visual part was initially created for Czech and later adopted to Russian, because of the fact that both of the languages belong to Slavic group of languages.

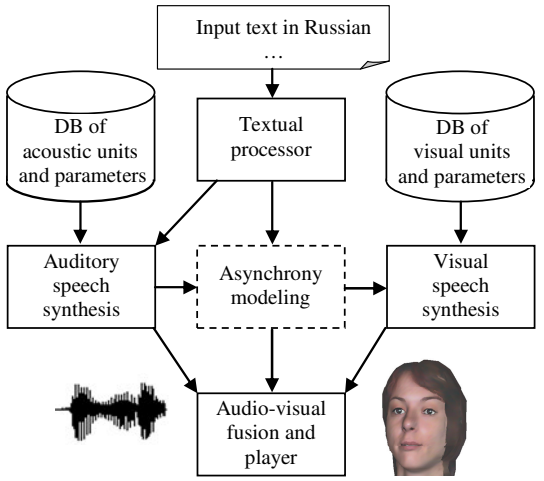

Figure 1: *General architecture of the bimodal synthesizer.*

3.1. Auditory speech synthesis

The applied TTS-synthesis model is based on allophones and multi-allophones natural waves (ANWs and MANWs) concatenation [7]. The speech prosody synthesis uses an original Accentual Units Portrait (AUP) model for a stylization of tonal, rhythmical and dynamic contours of a phrase. Fusion of these modules allows synthesizing speech with a high degree of intelligibility and naturalness. In the system an incoming orthographic text to be transformed into speech signal undergoes a number of successive operations carried out by specialized processors: textual, phonemic, prosodic and acoustical. The textual processor divides an orthographic text into utterances; transforms numbers and abbreviations into textual form; divides an utterance into phrases; places word stress (weak and strong); divides phrases into accentual units (AU), and finally marks the intonation type of the input phrase. The prosodically-marked text is sent to the phonemic processor, which performs the following tasks: phonemic transcription of the text; transforming the phonemic text into allophonic one; combining the allophones into multi-allophones. The prosodic processor calculates the target values of fundamental frequency (F_0), amplitude (A) and duration (T) for each allophone basing on the intonation type of the input phrase and extracting the corresponding Accent Units Portrait from the AUPs database. The acoustical processor uses output of the phonemic and prosodic processors to extract the appropriate allophones and multi-allophones natural waves from ANWs and MANWs database, modify their prosodic parameters and concatenate the derived ANWs and MANWs to continuous speech signal.

3.2. Visual speech synthesis

The audio-visual speech synthesizer is based on a parametrically controllable 3D model of a head. Movable parts are animated by a set of control points. The synthesis is concatenative, i.e. the descriptions of the visemes (in the form of the sets of control points) are concatenated to produce continuous stream of visual parameters. In the concatenative approach the co-articulation problem has to be solved to avoid unpleasant visual artifacts at the viseme borders. In our case the co-articulation is modeled by visual unit selection method [8] rather than with the help of dominance functions since the former method is able to better achieve articulatory targets important for visual perception of certain phonemes (for example, occlusions for /p/, /b/, /m/). The animation model uses the synthesized control points to move vertices of a 3D head model. For smooth movements of parts of a face the vertices surrounding the control points are interpolated. For acquisition of a static model of a head, 3D scanning technology is used. It employs a set of a camera, 4 mirrors and a dataprojector. The 3D information is composed from two viewpoints contained in one camera frame, where they are projected by the set of mirrors. The dataprojector generates vertical stripe light that moves over the scanned face horizontally during the scanning. As a result a 3D model of a face part of a specific person is obtained. Similarly the whole head model is acquired using the same system, while manually combining the views from different points.

3.3. Rule-based asynchrony modeling

Synchronization of virtual face and lip movements with synthesized acoustical signal is realized on the basis of information known about positions of beginning and end boundaries of each context-dependent phoneme (allophone) in the speech flow. Duration of every allophone is set by the auditory TTS system based on allophone average length and required speech tempo. To model bimodal asynchrony in the developed audio-driven talking head and take into account different speech rates the following 16 context-dependent timing rules for transitions between visemes were defined:

- from a pause to any vowel phoneme: from ¼ to ¾ of acoustical pause duration, but not more than 160 ms and not less than 80 ms;
- from a pause to any consonant phoneme: from ½ to end of pause, but not more than 120 ms and not less than 40 ms;
- from an unrounded vowel to a rounded vowel: from ½ to end of the first vowels, so the center of transition is approximately located at ¾ of the first vowel;
- from a velar consonant (/h/, /g/, /k/) to a rounded vowel: from the beginning to ½ of the consonant (consonant may disappear altogether);
- from a velar consonant to an unrounded vowel: from ¼ to ¾ of the consonant, so the center of transition should be placed in a middle of the consonant phoneme;

- from any dental or alveolar obstruent consonant to a rounded vowel: from ¼ to ¾ of the consonant;
- from any dental or alveolar obstruent consonant to an unrounded vowel: from ½ to the end of the consonant;
- from a consonant to a labial obstruent consonant (/v/, /f/, /p/, /b/): from ¼ to ¾ of the first consonant;
- from a vowel to a labial obstruent consonant: from ½ to end of the vowel;
- from a vowel phoneme or labial consonant to pause: from ¾ of the first phoneme to the end of the phoneme plus ¼ of its duration;
- from any obstruent consonant to a labial obstruent consonant: from ¼ to ¾ of the first phoneme;
- all other transitions between visemes are performed within the time from $^2/_3$ to the end of the first phoneme;
- for any second phoneme in a phrase or a syntagma the delay is additionally increased by ¼ relative duration of a preceding phoneme, but not more than 40 ms;
- for any third phoneme in a phrase the delay is additionally increased by $^1/_8$ relative duration of a preceding phoneme, but not more than 20 ms;
- for every stressed vowel all timings correspond to unstressed versions but with a shift of the viseme tail boundary ahead by $^1/_8$ relative duration of the following phoneme (excluding pause), but not more than 20 ms.
- for the stressed rounded vowels timings correspond to unstressed variants with an additional lag for the phoneme by $^1/_8$ relative duration of the preceding phoneme.

4. Cognitive experiments with talking head

The above-listed timing rules were implemented in the proposed 3D model-based AVSS. The cognitive experiments with the talking head consisted of two coherent parts: (1) evaluation of different kinds of asynchrony models of speech synthesizer aimed at the estimation of synthesized speech naturalness; (2) analysis and evaluation of speech intelligibility in noises. Three kinds of stimuli were applied at the speech perception experiments: (1) auditory synthesized speech; (2) audio-visual synthesized speech with the talking head; (3) pre-recorded auditory real speech (the same speaker's female voice was used for creation of the synthesized voice). Totally 20 phonetically-balanced sentences were selected before the testing and presented in a random order to each informant during the experiment. Each phrase is composed of 3-5 well-known meaningful Russian words connected by prepositions. However, all the phrases are meaningless on the whole or have a partial meaning so that to test human's visual and hearing perception only without a-priory semantic knowledge.

At the first step informants were asked to listen two times to a synthesized phrase (in order to avoid an effect of suddenness); after that they had to write down a chain of words they recognized. Then the subjects could perceive the same utterance said by the fully-functional talking head, at the second step the subjects had to examine four kinds of audio-visual synchronization models: (1) the baseline system without any asynchrony; (2) the talking head with the proposed asynchrony model; (3) a simple asynchrony model where a stationary delay of 150 ms is applied to the audio signal relatively to the corresponding video signal (V150A model means the video signal leads in 150 ms); (4) another simple asynchrony model where a stationary delay of 150 ms is applied to the video signal (A150V model). At this step the informants were asked to test the talking head and to evaluate the quality and naturalness of audio-visual synchronization ("in-sync" or "not in-sync") of the synthesized speech by a 5-point scale. The informants wrote down a sentence they recognized. At the last step they were asked to listen twice to the same pre-recorded phrase, but said by the real speaker and fill in the third string of the questionnaire. Such a cycle with different test phrases was repeated 20 times per each tester. Moreover, an additive acoustic noise of two types (bubble or white noise) with different intensity (SNR varied from 5 to 25 dB) was randomly introduced into the clean speech signal.

10 volunteers of 20-35 years old with normal hearing and eyesight took part in the experiment. Figure 2 shows a distribution of user evaluations averaged over all test phrases for each tester. All the persons identified mis-synchronization of auditory and visual speech cues for A150V model; two subjects of them did not perceive difference in synchronization for baseline talking head, V150A model and the proposed asynchrony model; two other persons did not find any difference in synchronization for V150A model and new asynchrony model; the remaining subjects confirmed that they see distinctions in the talking head with different asynchrony models. Surprisingly enough, but most of the respondents evaluated the baseline model with rather low marks, it was placed third in "the contest"; the majority of users gave the first place to the proposed original asynchrony model and only one person preferred V150A model. The informants were more tolerant to video signal leading than to audio signal leading.

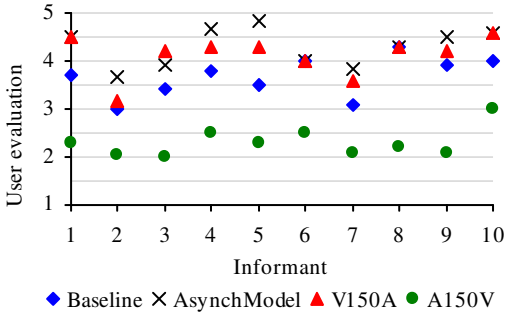

Figure 2: *Averaged user evaluations of asynchrony models.*

Figure 3 presents a distribution of user evaluations for four asynchrony models averaged over all the speakers in dependence of the signal-to-noise ratio (SNR) of bubble and white noise. It can be noted that with decreasing SNR informant's evaluations decrease as well, and this is true for every model excluding A150V. Important issue is the distance between evaluation marks for each pair of 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. Probably it could be caused by difficulties in the detection of starting and ending moments of speech signals. However, an advantage in naturalness of the proposed asynchrony model in noiseless conditions was appreciated by most of the informants.

These results coincide well with other recent cognitive studies [9, 10] that aimed at investigating the influence of stationary shifts of auditory or visual speech signals on the intelligibility of real audio and/or visual recordings. For example, in [9] it was shown that high speech intelligibility is preserved while visual speech leads up to 200 ms or audio speech leading is less than 30 ms. Work [10] states that the best speech intelligibility is reached with a stationary audio delay of 40-200 ms, and this is better than without any asynchrony. However, all these research studied stationary signal delays only, in the present investigation we analyze a new rule-based asynchrony model, which proposes humanlike dynamical asynchrony between flows of visemes and phonemes that makes audio-visual speech more expressive and natural.

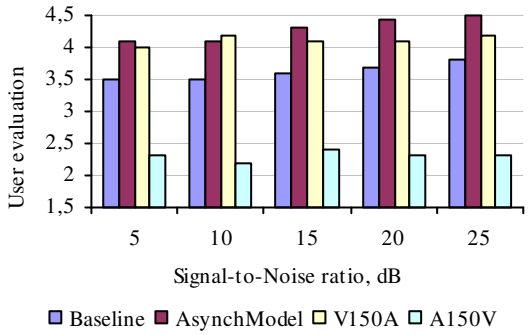

Figure 3: *Distribution of informant's evaluations vs. SNR.*

Figure 4 shows the results of the experiments on speech intelligibility. Rather low intelligibility values for synthetic speech could be explained by the inflective nature of Russian. Each stem corresponds with many endings, which are usually pronounced in continuous speech not so clearly as the beginning parts of words. Only grammatical rules and experience can help us to choose proper word-forms in a phrase. Phrases proposed to informants are grammatically correct but meaningless, so it was an additional difficulty for them to recognize proper word-forms from acoustic/visual information only. Figure 4 proves visual speech cue assists to recognize uttered speech better especially in noisy conditions. One can see that the distance between intelligibility functions grows while decreasing SNR value. In our experiments the AVSS system has outperformed the unimodal TTS by 6% on average. McGurk effect was repeatedly observed, for instance, in pairs of Russian words /oda/ (ode)- /oba/ (both), which were correctly recognized when adding visual speech only.

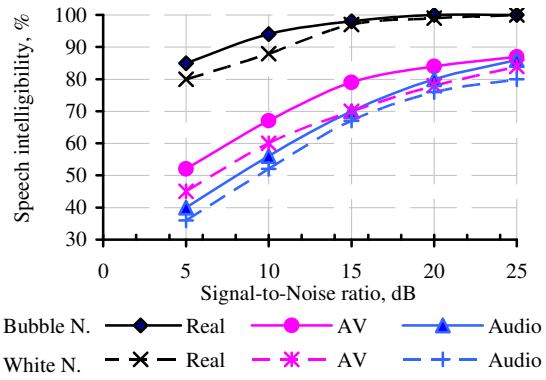

Figure 4: The *speech intelligibility with additive bubble noise (solid lines) and white noise (dashed lines).*

The intelligibility results are different for additive bubble and white noises. In white noise conditions the speech intelligibility is slightly lower than for bubble noise. This could be explained: white noise is more disturbing at listening whereas bubble noise is a more "pleasant" sound for ears. In bubble noise-added speech it is more difficult to detect speech boundaries, but speech is intelligible. However, white noise-added speech is problematic in general which influences intelligibility negatively. Statistically significant differences in user evaluation marks of asynchrony models were not observed in the experiments with bubble and white noises.

5. Conclusions

The proposed model of bimodal asynchrony was quantitatively compared with other models: the baseline talking head without any asynchrony model, V150A and A150V asynchrony models with stationary shifts of the signals. The results of the cognitive experiments have proved that a proper asynchrony modeling provides improvement of the user's perception of synthesized speech which is especially effective in clean speech conditions. It was discovered that it is better for human's perception to apply a simple asynchrony model with a stationary audio delay shorter than 200 ms rather than using a system, where modalities are regarded as completely synchronous.

The phenomenon of modalities asynchrony was studied with application to the Russian speech and language only. Russian has moderate (co-)articulation that is caused by cultural peculiarities and, probably, by the northern climate when people are used to speaking "through the lips" without a wide opening of the mouth. In contrast to speech in Russian, speaking in English (for native speakers) and in some other languages are characterized by hyper-articulation that results in greater asynchronies between both speech cues and other sets of context-dependent timing rules are needed for them. Comparison of bimodal speech asynchrony for Russian and English is being planned for further research.

6. Acknowledgements

This research was supported by the Ministry of Education of the Czech Rep., project No. ME08106; by the Grant Agency of the Czech Rep., project No. GAČR 102/09/P609; by the Belarussian Republican Foundation for Fundamental Research, project No. F08P-016; as well as by the Russian Foundation for Basic Research, projects No. 08-07-90002-Bel, 09-07-91220-CT and 08-08-00128.

7. References

[1] Sekiyama, K., "Differences in auditory-visual speech perception between Japanese and America: McGurk effect as a function of incompatibility", Journal of the Acoustical Society of Japan, 15:143-158, 1994.

[2] Nefian, A. et al. "A coupled HMM for audio-visual speech recognition", Proc. of International Conference on Acoustics Speech and Signal Processing ICASSP, pp. 2013-2016, 2002.

[3] Mattheyses, W., Latacz, L., Verhelst, W., and Sahli, H., "Multimodal Unit Selection for 2D Audiovisual Text-to-Speech Synthesis" Proc. of 5-th Joint Workshop on Machine Learning and Multimodal Interaction, Utrecht, The Netherlands, 2008.

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

[5] Govokhina, O., "Modèles de génération de trajectoires pour l'animation de visages parlants", PhD thesis, 2008.

[6] Karpov, A., Lobanov, B., Ronzhin, A., and Tsirulnik, L., "Audio-Visual Russian Speech Recognition and Synthesis for a Multimodal Information Kiosk", Proc. of 5-th International Conference on Neural Networks and Artificial Intelligence ICNNAI, Minsk, Belarus, pp. 81-86, 2008.

[7] Lobanov, B., and Tsirulnik, L., "Development of multi-voice and multi-language TTS synthesizer", Proc. of 11-th International Conference on Speech and Computer SPECOM, St. Petersburg, Russia, pp. 274-283, 2006.

[8] Krňoul, Z., Železný, M., Müller, L., and Kanis, J., "Training of coarticulation models using dominance functions and visual unit selection methods for audio-visual speech synthesis", Proc. of Interspeech'2006, Pittsburgh, PA, USA, pp. 585-588, 2006.

[9] Conrey, B., and Pisoni, D., "Audiovisual asynchrony detection for speech and nonspeech signals", Proc. of International Conference on Audio-Visual Speech Processing AVSP, St. Jorioz, France, pp. 25-30, 2003.

[10] Grant, K., and Greenberg, S., "Speech Intelligibility Derived from Asynchronous Processing of Auditory-Visual Information", Proc. of International Conference on Auditory-Visual Speech Processing AVSP, Aalborg, Denmark, pp. 132-137, 2001.