Virtual Speaking Head for Hearing Impaired Students

Eva Pajorová, and Ladislav Hluchý

***Abstract*—**Developed tool is one of system tools for easier access to various scientific areas and real time interactive learning between lecturer and for hearing impaired students. There is no demand for the lecturer to know Sign Language (SL). Instead, the new software tools will perform the translation of the regular speech into SL, after which it will be transferred to the student. On the other side, the questions of the student (in SL) will be translated and transferred to the lecturer in text or speech. One of those tools is presented tool. It's too for developing the correct Speech Visemes as a root of total communication method for hearing impared students.

***Keywords*—**Impared people, sing language, communication methods.

I. Introduction

THERE are also methods and systems for the SL recognition [14 -17], developed especially for Australian [18] and American [19] English. Since the SL is different in every country, it is necessary to develop special tools for the SL recognition, using special dictionaries. These projects are also aimed at general everyday communications.

Our developed tool is one of the actually developed software tools which ensure smooth link between regular distance learning and training of hearing impaired. For this, additional Sing Language (SL) information will be inserted in the main data stream of the video lessons. The corresponding SL information will be represented by video sequences with two SL interpreters; one will be contour image and second will be virtual speaking head. Both will be visualized scaled down in one of the lower corners of the main image. In order to significantly reduce this additional information, the contour images will be used instead of the full video. They are obtained after processing of the consecutive TV frames of the SL interpretation. The contour images represent very well the movements of the interpreter's hands and give very good vision of his/her face expression, which is of high importance for the sign comprehensibility. This approach permits hearing impaired to use the existing lessons for distance learning without delays, but together with the additional information in their mother language.

II. Related Approach

Second interpreter scaled down in one of the lower corners of the main image will be virtual speaking head, developed by our presented tool. Paper describes the tool for developing the second interpreter.

Lots of deaf hearing impared people are using lips reading as a main communication form. A viseme is a representational unit used to classify speech sounds in the visual domain. A "viseme" describes the particular facial and oral positions and movements that occur alongside the voicing of phonemes. Design tool for creating correct speech visemes is designed. Its composed from 5 modules; modul for creating phonemes, modul for creating 3D speech visemes, modul for facial expression and modul for synchronization between phonemes and visemes and last one modut to generate speech triphones. We are testing correctness of visemas on slovak speech domens. Up to now the regular Slovak speech visemes have not been developed. The paper describes developed tool.

The reason to develop our tool come out from requirements of deaf people. Deaf people comunication analyse looks that 80% of deaf people have interest to understand normal speeking speech. Deaf people community want to live between normal speaking people. They prefer oral communication before sign language. Over the years there have been many debates and studies done on how to communicate and educate the Deaf and Hard-of-hearing child. At first, people believed that Deaf children were [incapable of learning] and so didn't bother trying to communicate or educate them. But, over the years it was proved that Deaf children were [capable] and could learn to communicate and wanted to communicate, just like other people. Today, the issue has become; what is the best way to try and educate and communicate the Deaf [12], [13].

There are three main methods that have been developed and are used:

- Oral method (oralism)
- Manual method (manualism)
- Total Communication method

Each of these methods had various pros and cons; all should be carefully examined by the parents of a deaf child. The ultimate method for communication should be chosen based on how the child can be empowered and function in society, sadly hearing people don't always consider what is best for the Deaf.

The Oral method is a method for communication and educating deaf and Hard-of-hearing children using only the spoken language, lip reading, and voice training.The Manual method or manualism is based totally on Sign language and

E. Pajorová is with the Institute of Informatics, Slovak Academy of Sciences, Dubravska 9, 84507 Bratislava, Slovakia (e-mail: eva.pajorova@ savba.sk).

L. Hluchý, is with the Institute of Informatics, Slovak Academy of Sciences, Dubravska 9, 84507 Bratislava, Slovakia (e-mail: Ladislav.Hluchy@savba.sk).

using the hands or physical ways to communicate. The goal of this method is to provide a way for Deaf people to interact with others without the use of spoken language. Children taught using this method don't need to worry about learning to speak or develop lip reading skills. This method is considered the natural way that deaf children learn to communicate. Furthermore, it encourages a sense of pride in being Deaf.

Total Communication is a fairly new method for educating and communicating with Deaf and Hard-of-hearing children. The goal of this method is to *incorporate lip-reading, speech*, and Sign language so that a child can communicate effectively in almost any setting. Children taught using this method are given an opportunity to develop their voice as much as possible, as well as allowed to use the more natural, manual/visual way of communicating. The results are amazing, since children are taught how to interact in both the hearing and the Deaf world. Total Communication works well to educate Deaf and Hard-of-hearing children, since it doesn't matter if a child has a mild hearing loss or is profoundly deaf.This method works with the child, the parents, and the educators, so that everyone can learn and communicate to the best of their ability. Children are allowed to be themselves and learn in a way that suits their needs. Total Communication includes everyone -- the hearing world and the Deaf world.

So, what is the best way to try and educate and communicate the Deaf? The three main methods; the Oral method (oralism), the Manual method (manualism), and Total Communication.

After analyse we prefer total Communication which to incorporate *lip-reading*, speech and Sign language and it means that the dead child can communicate effectively.

The main goal of our work is to develop correct *visemes for lip-reading***.** We are testing the results on virtual slovak speeking head. We are using the triphonebased approach.

For generating of new sentences, we use a triphonebased approach [6]. Triphones are short pieces of motion sequence that span three phonemes, so each viseme is stored with its context and therefore captures all of the coarticulation effect caused by the direct neighbors. Our similarity measure is easily extended from visemes to triphones, and we can thus find the best overlapping triphone sequences in our database that match any new sentences that needs to be synthesized. Our work is based on dense 3D surface scans, which makes it more versatile than image-based techniques [6].

Facial animation is facing three different challenges:

- producing corect and realistic face shapes in every single frame of the animation
- creating a dynamically realistic face motion over time
- creating corect and realistic lip-speech animation

Lot of models [4,5,6,7] may be based on marker point positions, 3D scans or images. This approach facing the problem of defining how the parameters of the model vary over time. For speech synthesis, this involves the problem of coarticulation. Consecutive new approach[8] define dominance functions of phonemes that control the interaction between subsequent phonemes as applied to muscle-based systems [1]. Same systems are based on Hidden Markov Model [9] to learn the dynamics of speech from audio, and transfer this information to a face model. Another approach [10]uses regularization techniques to compute smooth curves for the model parameters over time. In this model, coarticulation is due to the smoothness of the curve and a statistical representation of the variance of each viseme. Instead of synthesizing motion entirely. New approach which bring novelty was build on Video Rewrite[11] stores triphone motions in a database, and stitches them together to produce new utterances.

Our goal is use the last one soffisticate Video Rewrite approach and arrange the best work bench for developing real speeking head in Slovak language. Video Rewrite method deduce a phoneme similarity measure. This quantitative similarity measure relaxes the selection rule of viseme grouping and offers further substitution options. Similarly to Video Rewrite, the optimal triphone sequence for the synthetic animation is found by minimizing an error function that takes both, viseme similarity and smoothness of the animation into account.

In our work, we use Lipsynctool for a statistical analysis of the distribution of mouth configurations in high-dimensional face space. The Lipsync Tool is an interactive application that allows users to create lipsync animation from audio files. The lipsynctool is based on laser light engines (LLE). This method estimates a low-dimensional, nonlinear manifold from a set of data points. In Lipsynctool system, LLE allows us to derive a highly specific criterion for viseme similarity that dictates appropriate triphone substitutions. LLE has been used previously by Wang et al.[28] as a representation that allows the separation of expression style from expression content with a bilinear model. Using a closely related Isomap method, Deng and Neumann[11] present a data-driven approach to speech animation where users can edit facial expressions in sequences.

In our work, we proposed the second interpreter for impaed people - the virtual slovak speech speeking head. Our novel selection method takes full advantage of the dual association between phonemes and visemes: not only can a phoneme take the visual appearance of several visemes, but visemes can be attributed to different phonemes as well. Our method determines visemes that can be used as a valid substitution for aspecific phoneme event.

III. Tool for Creating Correct Slovak Speeech Visemes

Tool for creating Slovak speech visemes is composed from 5 modules. Modul for creating phonemes, modul for creating 3D slovak speech visemes, modul for facial expression and modul for synchronization between phonemes and visemes and last one modut to generate slovak speech triphones (Fig. 1 shows schema). Each of modules can be as a separate modul or it can be include to the framework. For the synchronization all modules we use Lipsynctool.

Lipsynctool is able to synchronise phonemes together with creating own character visemes and with facial expresion modul as final triphones are developed with triphone muduls which are animate on the self developed face models.

All modules are creating tool for developing slovak speech speaking heat.Virtual slovak speeking head is able navigate and instruct the crisis situations as is evacuation the people from big halls, also to evacuation shopping centre and its could be usefull in crisis situation as are natural disasters (fires floods...).

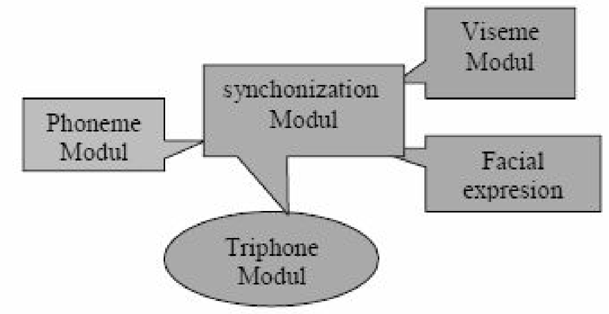

Fig. 1 Schema of modules of tool for Slovak speech visemes

A. Phoneme Modul

As a Slovak speech phonemes input we use phonemes which have been generated by slovak speech synthesis, developed in our institute [1],[2],[3]. In a scope of the activites was developet lot of slovak speech results as is for example the text collection Slovak National Corpus. Other activity is focused on speech database building for speech recognition purposes. It was developed and build a one speaker speech database in Slovak for experiments and application building in unit-selection speech synthesis. To build such a database, it was to exploit as much of the existing speech resources in Slovak as possible, and to utilize the knowledge from previous projects and to use the existing routines developed at our department. As a input file was use audio file (.wav) and slovak text.

B. Viseme Modul

The viseme targets for us character should be somewhat exaggerated from real mouth positions. We use frames from laser scanner see Fig. 2. The mouth positions can be approximately grouped into vowels and consonants. Vowels are the voiced sounds. At their peak position, they will have a comparitively open position to its neighboring consonants. Realism is effected by three factors in the viseme representation:

The width/backness of the mouth.

The openness of the mouth.

The speech functionality generates articulation information that controls the openness and emphasis of a phoneme (per frame - phn_vis) or on a curve (phn_env). Mouths should be created as if they were being emphasized. The speech system will generate information to control "how much" of the phoneme is turned on. The speech system won't exaggerate the phoneme unless the speech is emphasized.That said, certain phonemes are more open than others by default.

The rounded-forwardness of the mouth.

For good lipsync, it is necessary to try to capture the various curved phonemes. Each phoneme will be a combination of those attributes, most are a combination of 2, openness and width-backness. or openness and rounded-forwardness.

Referring the to viseme images, we can classify each phoneme as the contribution of a width-backness factor, the openness factor, and the rounded-forward factor. This is subjective, and there may be disagreement. The viseme is listed following by a contributions. If the contribution is zero, this means the "neutral position" for forward-rounded, width-backness. For openness, zero means closed.

C. Modul for Facial Expression

Expression is a group of facial parameter values that together transform the neutral face into an expressive face [2]. Expressions simulation and animation can be divided into two groups.The first one are animated or recorded visemes (and corresponding phonemes), which act on the region of the lips and the mouth and form segments of words. The second one - emotions act on any part of the face. Contrary to other regions of human face, the lips are mainly characterized by their contours. Several models of lips were created. One of them uses algebraic equations for contours approximation. From a multidimensional analysis of a real speaker's gestures, visemes and corresponding phonemes can be extracted. After lips, the most visible moving part of the face by speaking is jaw. Jaw kinematics can be modeled manually or automatically from the data recorded by video, mechanical or opto-electronic sensors. In realistic image synthesis, superpositions of jaw transformations and lips shapes must be integrated with emotions. Primary emotions are for example surprise, fear, anger, happiness, sadness. Basic expressions and their variants can be defined using snapshots of the computer or real model of the actor [3].

D. Synchronization Modul

Phonemes are expressed using a combination of jaw, tongue, and lip movements. Each of these parts of the mouth is called a channel in the Speech action clip. Having all three channels' animation within one clip helps you create smooth blending between the visemes because you can control the lip-sync animation as one entity. Within the clip, however, the animation of the lip, jaw, and tongue controls are separated into different curves so that you can control each channel on its own. Annosoft Lipsync automatically determines how open the mouth should be, and when. The viseme should try to be as accurate a representation as possible, obviously. To get there, we need to subdivide the vowels so that we can make an accurate mouth reference.

E. Triphone Modul

The In linquistics a triphone is a sequence of three phonemes. Triphones are useful in models of natural language processing where they are used to establish the various contexts in which a phoneme can occur in a particular. Lipsynctool allows us to define an arbitrary number of visemes. The Lipsync Tool is application providing powerful automatic lipsync synchronization and timeline editing capabilities

IV. Conclusion

The goal of our work is to develop such sophisticated tool, which will be providing the regular Slovak speaking head. We are testing the Slovak visemes in Lip-reading method for Slovak deaf people. Currently we try to include them to the Total communication method. Lip-reading together with Sing language could be help the Slovak deaf people. The most popular systems used to transfer visual information for distance training for hearing-impaired people via Internet, which is based on the video compression standards H.261 and MPEG-4. In this case, the basic information is represented with video sequences, containing the image of the Sign Language interpreter [1].

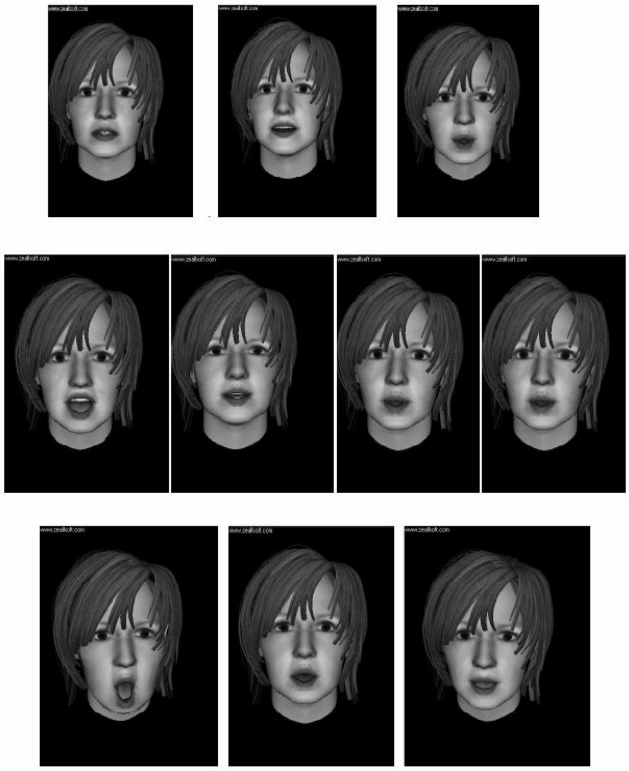

Fig. 2 Examples of 3D Slovak speech visemas

Acknowledgment

This work was supported by VEGA - research project 2/0054/12 and by Slovak Research project CRISIS - ITMS - 26240220060 - Research and development of new information technologies for prediction and solution of crisis situations and security of habitants.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Hrúz M., Krňoul Z., Campr P., and Muller Ludek M.S.: Tovards Automatic Annotation of Sing Language Dictionarz Corpora, 2011 TDS Pilsen, page 331 - 339.

[3] May, Qin Caiy, David Gallupz, Cha Zhangy, and Zhengyou Zhang: 3D Deformable Face Tracking with aCommodity Depth Camera. http://research.microsoft.com/en-us/um/people/qincai/papers/eccv2010.pdf

[4] Peter Drahoš and Martin Šperka, Face Expressions Animation in e-Learning. lconf06.dei.uc.pt/pdfs/paper14.pdf.

[5] Albrecht, J. Haber, and H.-P. Seidel. Speech Synchronization for Physicsbased Facial Animation. In V. Skala, editor, Proc. 10th Int. Conf. on Computer Graphics, Visualization and Computer Vision (WSCG 2002), pages 9–16.

[6] ApJ. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate visible speech synthesis based on concatenating variable length motion capture data. IEEE

[7] A. Wang, M. Emmi, and P. Faloutsos. Assembling an expressive facial animation system. In Sandbox ’07: Proceedings of the 2007 ACM SIGGRAPH symposium on Video games, pages 21–26. ACM Press, 2007.

[8] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. In SIGGRAPH ’02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques, pages 388–398. ACM Press, 2002.

[9] M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. In N. Magnenat Thalmann and D. Thalmann, editors, Models and Techniques in Computer Animation, pages 139–156. Springer, Tokyo, 1994.

[10] M. Brand. Voice puppetry. In SIGGRAPH ’99: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pages 21– 28. ACM Press/Addison-Wesley Publishing Co., 1999.

[11] I.-J. Kim and H.-S. Ko. 3d lip-synch generation with data-faithful machine learning. In Computer Graphics Forum, Vol. 26, No. 3 EUROGRAPHICS 2007, 2007.

[12] C. Bregler, M. Covell, and M. Slaney. Video rewrite:driving visual speech with audio. In Computer Graphics Proc. SIGGRAPH’97, pages 67–74, 1997.

[13] Meadow – Orlans, Kathryn P., Mertens, Donna M., & Sass – Lehrer Marilyn. (2003) Parents and their Deaf Children, Washington D.C.; Gallaudet University Press.

[14] D C. Sylvie, W. Ong, S. Ranganath. Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning, IEEE Trans. on Pattern Analysis and Machine Intelligence 27(6):873-891, 2005.

[15] R. Bowden, D. Windridge, T. Kadir, A. Zinaerman, M. Brady. A linguistic feature vector for the visual interpretation of SL. Proc. of 8th EU Conf. on Computer Vision 2:391–401, 2004.

[16] A. Edwards. Progress in Sign Language Recognition, Proc. Gesture Workshop, pp. 13-21, 1997.

[17] B. Bauer, K. Kraiss. Towards a 3rd Generation Mobile Telecommunication for Deaf People, Proc. 10th Aachen Symp. Signal Theory Algorithms and Software for Mobile Comm., pp. 101-106, 2001.

[18] E. Holden, G. Lee, R. Owens. Australian SL recognition, Machine Vision and Applications, pp. 312-320, 2005.

[19] T. Starner, A. Pentland. Visual recognition of American Sign language using hidden Markov models. Proc. of the Intern. Workshop on Automatic Face- and Gesture-Recognition, pp. 189–194, 1995.