Towards a Personalized Multimodal System for Natural Voice Reproduction

Stelios Gavras*†, Spiros Baxevanakis*‡, Dimitris Kikidis¶, Efthymios Kyrodimos¶ and Themis Exarchos*§
*Department of Informatics, Ionian University, Corfu, Greece
†Email: stelgavras@gmail.com
‡Email: sprbax@gmail.com
§Email: exarchos@ionio.gr Telephone: +30 2661087855
¶Voice clinic, Medical School, National and Kapodistrian University of Athens, Athens, Greece

***Abstract*—Loss of voice, which may be due to problems in various organs of the human voice system is a major disability which most of the time results in social isolation. In most cases of complete voice loss, the problem cannot be improved over time and despite the fact that methods have been developed to reconstruct voice, none come close to the natural pitch of the person's voice. The present work proposes a novel personalized system that integrates audiovisual information and autonomous learning in order to reconstruct a disabled person's voice. The system makes use of a smartphone's camera and microphone while the majority of computations will not be performed locally but in the cloud. Using computational methods for signal processing as well as machine learning algorithms we believe that, a voice similar to the user's own voice can be reproduced. Finally the aforementioned technologies will be incorporated into a personalized multimedia interface and the system will be evaluated during a multi-phase period by patients with serious chronic voice problems.**

I. Introduction

Voice is one of the main characteristics of the human race. It is essential in means of communication and expression of thoughts and feelings and is essential for the development of civilization. Larynx is an important organ as it ensures breathing and speech. It is the narrowest point of the airway tract and thus is a point of variance of interventions in order to maintain breathing function and consequently survival. Voice production is performed by tuning the vocal chords during exhalation. Each person has a separate voice tag, unique, recognizable and part of his or hers personality. Loss of voice, which may be due to problems in various organs of the human voice system and consequently causes significant communication difficulty, is a major disability. It affects a person's ability to communicate, to express feelings, affects his/hers job, daily life and relationship with relatives and significant others. The increase in the average age and the large number of people with severe disabilities, such as blunt trauma after a car accident, have contributed to the growth of population that experiences difficulties in voice communication, since it is necessary for those to carry out permanent tracheostomy. In the US, number of patients who undergo a tracheostomy and consequently lose their speech ability is over 100.00 per year [1]. The increase of hospitalization rate in intensive care units has increased by 11% over the last 10 years [2] and similarly the number of patients at risk of speech loss has increased, because 40% of the hospitalized patients are tracheostomy candidates due to prolonged intubation. In particular for Greece, the increased smoking rate (more than double of the European average) leads to an increase in total laryngectomy due to laryngeal cancer (95% in smokers [3]). Trends of laryngeal cancer are the highest in European level, whereas it ranks first among the malignancies . Total or significant loss of voice, may also be caused by various other diseases and conditions: (i) Laryngeal or thyroid diseases, (ii) Paralysis of vocal cords, (iii) Respiratory diseases, such as chronic obstructive pulmonary disease, (iv) progressive aphasia (dementia), (v) Other neurological disorders which affect laryngeal muscles mobility and coordination (e.g., myasthenia gravis, multiple sclerosis, Parkinson's disease), (vi) Damage of the laryngeal nerve after surgery, especially thyroidectomy. As a result, individuals with partial or total loss of voice are socially isolated, resulting in other secondary disorders such as depression and impairment of cognitive functions, which in turn worsens their quality of life. In most cases of complete loss of voice, the problem cannot be improved over time, and despite the fact that many methods have been developed to restore voice, they have a poor voice effect, while at the same time there is no possibility of speaking with the natural pitch of the voice of the person.

The solutions currently available for the partial restoration of speech are unsatisfactory both in terms of quality and understanding, as well as, in the degree of patient adoption. The development of esophageal speech (speech from the stomach) is achieved by few patients but has an unacceptable voice effect. Esophageal speech is the hardest vocal rehabilitation method to learn. It needs a long period of time to practice and requires the patient to be in good physical condition. The use of a special device (laryngophone) is not very useful, as the voice produced is mechanical and sometimes obscure. Therefore, it has been more or less abandoned. Finally, the placement of a vocal prosthesis with systolic development between the trachea and the esophagus is a surgical method of rehabilitation with high chance of failure, risk of infections, permanent colonization of the prosthesis by biofilms (as a foreign body) and other complications, while requiring patient education

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 579, Proposal's Title: Application of multimodal interfaces for people with loss of voice for speech reproduction with natural voice, Acronym: Let's Talk!)

978-1-7281-5919-5/20/$31.00

and continuous care [4]. Also, its cost is relatively large, requiring a six-month repositioning, creating additional costs, need for hospitalization and health services overload, since repositioning requires admission, whereas numerous technical problems cause repetitive visits.

Our proposed framework aims to create an innovative personalized multimodal interface, which will be provided to users through an application, based on cloud computing infrastructure (for smart phones and / or tablets) in order to allow them to communicate with their natural voice tone, as it was before the partial or total loss of their voice. The multimodal interfaces will accept as input video from the user's lips as well as any sounds generated during the person's attempt to speak. Through the fusion of information from the lips (videos) and the sounds produced, the application, after appropriate adaptation to the user's profile and through the use of vocabularies (so that the words produced make sense and the sentences are logical), will produce in a written text the user's speech. Then, using existing tools (Praat, Mary, Google TTS) and availability of user's voice recording prior to loss of voice (either with pre-recording, or through previously recorded voice/video), the voice effect will be generated. Figure 1 illustrates the general concept the proposed system.

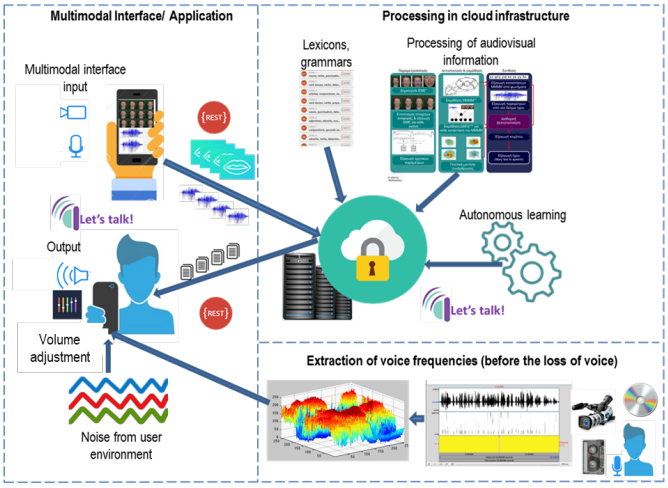

Fig. 1. An overall view of our proposed framework.

II. Related Work

Multimedia as a research topic is more than just a combination of diverse data modalities [5], i.e., audio, video, text, graphics, etc. Essentially, it is the amalgamation and the interaction among these divergent media types that lead to the formation of some extremely challenging research opportunities such as that of speech reading and reconstruction [6].

Traditionally, multimedia systems have been grouped into text, audio and video based information systems. One of the most fundamental problems in multimodality is that of speech reading and reconstruction systems. This problem belongs to one of those research arenas which are cross domain and exploit the full breadth and depth of multimedia research since it involves not just speechreading but the synchronization of video with lip movements as well as reconstruction of the audio. Speechreading involves looking, perceiving, and interpreting spoken symbols. Speech is considered to be composed by phonemes that are considered to be the smallest distinct and detectable units of sound in a given language [7]. Phonemes are produced by different movements of teeth, tongue, lips, etc. Visemes, on the other hand, are facial images including jaw and lip movements that are used to define and generate sound [8]. The task of speechreading is made difficult by the fact that often several phonemes correspond to a single viseme, thus producing ambiguity when trying to infer speech from visemes only. For example, it is very difficult to distinguish the English characters 'p' and 'b' just by looking at the image of a speaker who speaks both of them. That is because English characters like 'p' and 'b' belong to the same viseme class. Similar is the case with expressions like "Elephant Juice" and "I love you" which, though having similar visemic appearances, definitely have very different sounds and meanings [9]. In the recent years, Automatic Speech Recognition (ASR) systems have gained significant traction, with systems being deployed widely in cars, mobile phones, homes, [10], [11].

However, considering settings such as that of a car on a road, it is not ideal and is plagued by very low signal to noise ratio [12]. This leads to speech recognition systems failing utterly in the absence of usable audio [13]. However, these fallacies with ASR systems can be solved satisfactorily by deploying speech reading and reconstruction systems which can augment the understanding of ASR systems or can even reconstruct the speech for them. With the advent of multiple frontal cameras in embedded devices like mobile phones, webcams and cars, additional information in the form of visual output of speech such as lip movements is easily accessible. Currently, the visual feed thus obtained is being ignored or wasted. This can be effectively integrated with ASR system via the use of technology proposed in the current paper. Another crucial case is of venues requiring critical security infrastructure such as cockpits, battlefields and even public places such as railway stations, airports, etc. In this case, the cameras either do not record the audio or the audio recorded is too feeble or jagged to be of any use. This situation unequivocally calls for a solution. Additionally, considering security and crime scenarios themselves, due to the availability of camera footage, visual cues such as lip-movements have been used in the past for solving deleterious crimes [14]–[20]. However, professional lipreaders are required for rendering their services in these cases, which are not only expensive but highly limited. The single solution for all these challenges is a system having speech reading and reconstruction capabilities which can also effectively integrate with ASR systems. Some of the earliest works reported in the field of lipreading are those by Potamianos et al. [21] and Lan et al. [22]. Recent works such as those by Cornu and Milner [23] and Akbari et al. [24] perform the task of lipreading on the GRID database [25]. Previous works [23], [24], focused on single view based lipreading often with hand-picked features.

Neural networks [26], [27], have been used or a combination of neural networks with hidden Markov models [28], [29], as well as other algorithms [30], [31]. These approaches have achieved remarkable results in terms of accuracy, but have been evaluated in standard data sets rather than in everyday practice. Also, limited methods use grammars and vocabularies, but none has been applied to the Greek language, with the

exception of Google Text to Speech [32], which is applied for text to speech conversion rather than speech recognition from images. Also, none of the above methodologies offer a complete solution that results in talking with the user's natural voice before it was lost. The system proposed in this work aims to develop an integrated application that will use multi-modal information, its processing in mobile and cloud infrastructure, to reproduce the user's speech in real time, with the natural tone of his voice.

III. Main components

A. Multimodal interfaces and application architecture

The system that the authors propose, is based on the use of smartphones. Considering that they have the necessary hardware (camera, microphone, earphone and internet connection), is an advantage over other devices both because the ease of use, (the user is not forced to use a new device in everyday life) and cost. Through the built-in camera and microphone on the smartphone, the app will receive the input from the user. Part of the process (lip image segmentation) will be performed in the smartphone, so that the transmission of videos of large size to the cloud infrastructure is avoided. The segmented lip images will be sent through REST Web Services to a remote web server over secure encrypted HTTPs. The server will process the two signals, the voice and the segmented images. By combining them, it will extract the user's phrase into text and send it back to the application for playback, which is converted into speech (due to the small size of text files, this approach is preferable compared to sending audio files). The main processing will be performed in the cloud computing infrastructure and not in the application. The development methodology will be based on the analysis of requirements from the users and the experts.

B. Speech Recognition using audiovisual information

The method to be used for speech recognition from audio-visual information represents visual data using a compact set of parameters obtained from building an active appearance model (AAM) [33] on landmark-annotated facial images. Audio data is represented using both discrete phonemes and continuous speech features. We will then use a variable length Markov model (VLMM) [34]–[36] to learn a language model on phonetic data such that commonly occurring fragments of speech are automatically segmented as states. The audio and visual parameters corresponding to those states are extracted and modelled using a shared Gaussian process dynamical model (SGPDM) [37], [38]. This will result in a switching shared Gaussian process dynamical model (SSGPDM) with the learnt VLMM states as switching states. During synthesis, we will use the trained VLMM and the test phonetic streams to infer VLMM states such that the corresponding SGPDMs indexed by those states can be used to predict visual parameters from audio. We propose a synthesis algorithm which models both anticipatory and carry-over coarticulation. Figure 2 shows an overview of the proposed method. To further convert text into audio, tools such as Praat: doing phonetics by computer [39], MARY Text-to-Speech System [40] and Google Text to Speech [32] will be used. It is noteworthy that the volume of the speech output will be automatically calculated based on the user's ambient noise (from the mobile microphone).

C. Profiling and autonomous learning

The application will integrate an autonomous learning process that allows users to build their profiles easily and quickly. When the application is initiated, the user will provide information about his/hers demographic and social data, so that speech recognition algorithms know the context in advance (for example type of work environment, place where the patient lives, habits, behavior etc). Based on this context, vocabulary matching for the user will be easily achieved. In addition, users after each conversation through the multimodal interface will be able to evaluate the result through the application. This result will be used by autonomous learning algorithms so that each time the application is used, it can sequentially perceive the user's requirements. The techniques to be used include dynamically Bayesian Networks, which use prior conditions to predict-create the next ones.

D. Natural Language Processing and Voice Reproduction

To convert the text into natural audio speech, the tools Praat: doing phonetics by computer, MARY Text-to-Speech System and Google Text to Speech will be adapted as follows: (i) Extraction of natural user speech frequencies from pre-existing video or recorded speech before the loss of voice (Praat), (ii) Preprocessing of the text for its normalization (MARY), (iii) Natural language processing for the lexical analysis and annotation (MARY), (iv) the calculation of the acoustic parameters, translating the linguistically annotated symbolic structures into a table containing existing words and expressions (MARY), (v) the composition and production of this table to sound, using the natural voice of the user (MARY, Google TTS).

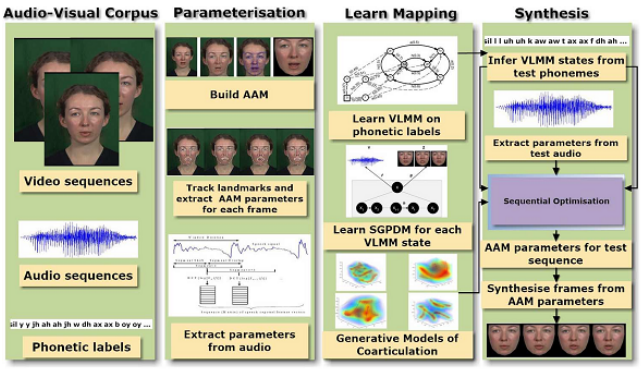

Fig. 2. The method for speech recognition using audiovisual information.

E. Evaluation

Patients with a significant degree of difficulty in speaking will participate in the study in two stages: stage A, will be a key pillar for the user requirements gathering process. Specifically, they will be involved in the design phase, with personal interviews by otolaryngologists and speech therapists, questionnaires and specially structured workgroups (patient focus groups, doctors and software engineers) from which they will collect their requirements. In stage B a pilot study will be carried out. Both stages of the study will be in accordance with the requirements and specifications of the European Union Clinical Trials Regulation (536/2014). All patients will be

informed in detail about their participation in the study and will sign a consent form. They will be also informed that they can leave the study at any time without this option to have any effect on their treatment. The application will allow for evaluation in 30 patients with chronic severe difficulty in vocalization. The underlying diseases will involve total laryngectomy due to cancer of the larynx, permanent tracheostomy due to injury or paralysis of the vocal cords and neurological diseases affecting the larynx. The pilot study (stage B) will be conducted in two phases. During the initial phase, which will take 3 months, material containing patient's voice will be chosen and patients will be educated to the application usage means of evaluation. The application will record the frequency and the duration of usage, the rate of correct word recognition by the application (with interaction with the patient in a controlled environment), and also the level of understanding the outcome from the clinicians and caregivers/relatives of the patients. After three months, customized questionnaires for these conditions will be formulated in order for patient and relatives to give an overall assessment and recommendations to improve the functionality of the platform. Upon completion of the first phase, which will involve 10 patients, the application will be updated based on the observations and feedback. The second phase of the evaluation will be conducted with the new version of the application, after integration of changes and modifications. The final assessment and evaluation of the application will be performed in the second phase, where the application will be tested in 20 patients.

IV. Discussion

The proposed approach is expected to contribute significantly in gaining know-how in the field of information technologies, multimodal interfaces, machine learning and data modeling. Patients who underwent permanent tracheostomy (as in laryngectomy) or had severe damage to the rib and therefore the mobility of the vocal cords or suffer from a loss of voice situations mentioned in the previous sections, will have an effective, easy to use and low cost application, that will help them to cope with their disability. Patients who are unable to express themselves experience severe deterioration not only to every day quality of life but to their mental and emotional status as well. Reliable and standardized expression via the proposed platform could be a great relief for their emotional status. The feeling that one can again be part of a group or at least be able to satisfy his needs via his voice is a major social impact factor that our proposal will deliver to these people in need. The opportunity for the disabled to express needs, feelings and ideas can offer them the freedom to interact with their relatives and their social environment in many levels.

References

[1] Cheung NH, Napolitano LM. Tracheostomy: epidemiology, indications, timing, technique, and outcomes. Respir Care. 2014 Jun;59(6):895-915

[2] Owings MF, Kozak LJ. Ambulatory and inpatient procedures in the United States,.National Center for Health Statistics, Vital Health Statistics. www.cdc.gov/nchs/data/series/sr_13/sr13_139.pdf.

[3] Kikidis D, et al Continuation of smoking after treatment of laryngeal cancer: an independent prognostic factor?. ORL J Otor. Relat Spec. 2012;74(5):250-4

[4] Elmiyeh B, et al. Surgical voice restoration after total laryngectomy: an overview. Indian JCancer. 2010 Jul-Sep;47(3):239-47.

[5] R. Shah and R. Zimmermann. 2017. Multimodal analysis of user-generated multimedia content. Springer.

[6] R. Shah, Y. Yu, and R. Zimmermann. 2014. Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In Proceedings of the 22nd ACM international conference on Multimedia. ACM, 607–616.

[7] S.E Shaywitz. 1996. Dyslexia. Scientific American 275, 5 (1996), 98–104.

[8] C. Benoit, T. Lallouache, T. Mohamadi, and C. Abry. 1992. A set of French visemes for visual speech synthesis, 485–501 pages.

[9] D. Jachimski, A. Czyzewski, and T. Ciszewski. 2017. A comparative study of English viseme recognition methods and algorithms. Multimedia Tools and Applications (2017), 1–38.

[10] J.R. Allen, D.M. West. 2018. How artificial intelligence is transforming the world. https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/

[11] Business Wire. 2018. European $1.66 Billion Speech and Voice Recognition Market Analysis 2016-2018 Forecast to 2025 - Key Players are Microsoft, Nuance Comms, and iFlytek - ResearchAndMarkets.com. https://www.businesswire.com/news/home/20180417005875/en/European-1.66-Billion-Speech-Voice-Recognition-Market

[12] R. Singh, R.M Stern, and B. Raj. 2002. Signal and feature compensation methods for robust speech recognition. Noise reduction in speech applications 7 (2002), 219.

[13] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong. 2015. Robust Automatic Speech Recognition: A Bridge to Practical Applications. Academic Press.

[14] BBC One Scotland. 2005. Frontline Scotland reveals new evidence in Arlene Fraser murder case. http://www.bbc.co.uk/pressoffice/pressreleases/stories/2005/10_october/19/fraser.shtml

[15] 2012. John Terry cleared of racism against Anton Ferdinand. (2012). http://www.bbc.com/news/uk-england-london-18827915

[16] 2012. John Terry defence 'improbable, implausible and contrived'. (2012). https://www.bbc.com/sport/football/19842795

[17] 2018. Judgment of John Terry. https://www.judiciary.gov.uk/judgments/r-v-john-terry-judgment/

[18] 2018. Nine jailed over Euro 11m stolen goods. (2018). http://news.bbc.co.uk/2/hi/uk_news/england/2806155.stm

[19] D. Conn. 203. John Terry judgment: Main findings of the FAś regulatory commission. (2003). https://www.theguardian.com/football/2012/oct/05/john-terry-judgment-commission

[20] A. Palmer. 2003. Lip reader saw Fraserś incriminating conversations. (2003). https://www.telegraph.co.uk/news/uknews/1420816/Lip-reader-saw-Frasers-incriminating-conversations.html

[21] G. Potamianos, C. Neti, J. Luettin, and I. Matthews. 2004. Audio-visual automatic speech recognition: An overview. Issues in visualand audio-visual speech processing 22 (2004), 23.

[22] Y. Lan, B.-J. Theobald, R. Harvey, Eng-Jon Ong, and Richard Bowden. 2010. Improving visual features for lip-reading. In Auditory-Visual Speech Processing 2010.

[23] T. Le Cornu and B. Milner. 2015. Reconstructing intelligible audio speech from visual speech features. In Sixteenth Annual Conference of the International Speech Communication Association.

[24] H. Akbari, H. Arora, L. Cao, and N. Mesgarani. 2017. Lip2AudSpec: Speech reconstruction from silent lip movements video. arXiv preprint arXiv:1710.09798 (2017).

[25] M. Cooke, J. Barker, S. Cunningham, and X. Shao. 2006. An audiovisual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America 120, 5 (2006), 2421–2424.

[26] Y. Mroueh, et al, Deep multimodal learning for audio-visual speech recognition. In 2015 IEEE (ICASSP), pages 2130–2134. IEEE, 2015.

[27] O. Koller, H. Ney, and R. Bowden. Deep learning of mouth shapes for sign language. Proc. of the IEEE Int. Conf. on Computer Vision Workshops, 85–91, 2015.

[28] G. Galatas, et al. Audio-visual speech recognition incorporating facial depth information captured by the kinect. In (EUSIPCO), 2012 2714–2717.

[29] K. Noda, et al,. Audio-visual speech recognition using deep learning. Applied Intelligence, 42(4):722–737, 2015.

[30] J. S. Chung and A. Zisserman. Lip reading in the wild. In Proc. ACCV, 2016.

[31] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016

[32] Google LLC, "Google Text-to-Speech", retrieved 10 May 2020 from https://play.google.com/store/apps/details?id=com.google.android.tts

[33] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models,"IEEE Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, 2001

[34] D. Ron, et al, "The power of amnesia: Learning probabilistic automata with variable memory length,"Mach. Learn., vol. 25, pp. 117–149, 1996.

[35] I. Guyon and F. Pereira, "Design of a linguistic postprocessor using variable memory length Markov models," inICDAR'95: 1995, pp. 454–457

[36] A. Galata, N. Johnson, and D. Hogg, "Learning variable length Markov models of behaviour,"Comput. Vision Image Under., vol. 81, no. 3, pp. 398–413, 2001.

[37] S. Deena and A. Galata, "Speech-driven facial animation using a shared Gaussian process latent variable model," inISVC'09: 2009.

[38] CH Ek, et al, "Gaussian process latent variable models for human pose estimation," Proc. 4th Int. Work. on Machine Learning for Multimodal Interaction, 2007.

[39] Boersma, Paul & Weenink, David (2020), "Praat: doing phonetics by computer [Computer program]". Version 6.1.14, retrieved 10 May 2020 from http://www.praat.org/

[40] "The MARY Text-to-Speech System (MaryTTS) [Computer Program]", retrieved 10 May 2020 from http://mary.dfki.de/