(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
02 February 2023 (02.02.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/007509 A1

(51) International Patent Classification:
G06V 40/16 (2022.01) *G10L 25/75* (2013.01)
G09B 19/04 (2006.01) *G09B 5/04* (2006.01)
G10L 21/0356 (2013.01) *G09B 19/06* (2006.01)
G10L 21/12 (2013.01) *G10L 21/0332* (2013.01)
G10L 21/18 (2013.01) *G10L 21/043* (2013.01)
G10L 25/72 (2013.01) *G10L 21/055* (2013.01)

(21) International Application Number:
PCT/IN2022/050658

(22) International Filing Date:
22 July 2022 (22.07.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
202121033606 27 July 2021 (27.07.2021) IN

(71) Applicants: **INDIAN INSTITUTE OF TECHNOLOGY BOMBAY** [IN/IN]; Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra 400076 (IN). **DIGITAL INDIA CORPORATION** [IN/IN]; 4th Floor, Samruddhi Venture Park, Central MIDC Road #2, Andheri (East), Mumbai, Maharashtra 400093 (IN).

(72) Inventors: **PANDEY, Prem Chand**; B-99, Lakeside White House, IIT Bombay, Powai, Mumbai, Maharashtra 400076 (IN). **KACHARE, Pramod Haribhau**; A11-201 Prakruti Palms, Brahmand, Dharmachapada, Thane, Maharashtra 400706 (IN). **DASGUPTA, Hirak**; Nabajiban Colony, PO Bisharpara, Birati, Kolkata, West Bengal 700051 (IN). **NATARAJ, Kathriki Shambulingappa**; House No. 774/31, Shivakumaraswamy Layout Stage 2, Davangere, Karnataka 577005 (IN). **MANE, Vishal**; B-1201 Florida Watercolor, Keshavnagar, Mundhwa, Pune, Maharashtra 411036 (IN).

(74) Agent: **ROY CHOWDHURY, Mahua**; 2A/54 Kalpataru Estate, JV Link Road, Andheri East, Mumbai, Maharashtra 400093 (IN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) Title: METHOD AND SYSTEM FOR TIME-SCALED AUDIOVISUAL FEEDBACK OF SPEECH PRODUCTION EFFORTS

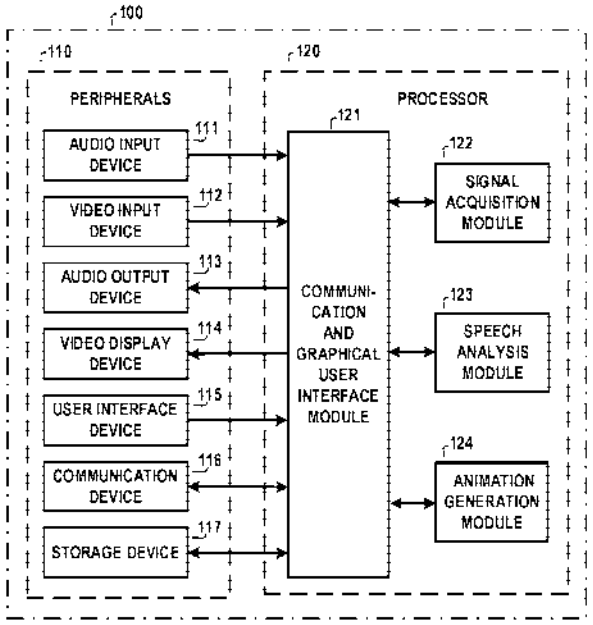

Figure 1

(57) Abstract: In the present disclosure, a method and a system are disclosed for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production of a plurality of persons. The system (100) comprises at least one processor (120) and a plurality of peripherals (110) interfaced with the processor (120). The processor (120) receives an audiovisual signal comprising each person's speech as the audio signal and the frontal view of each person's face as the video signal from the peripherals, processes the audiovisual signal, displays a time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face with a settable slowdown factor, and outputs a correspondingly time-scaled audio signal. The disclosed method and system may be useful in speech training aids for the hearing- impaired children and second-language learners, voice disorder diagnosis, and speech rehabilitation therapy.

[Continued on next page]

WO 2023/007509 A1

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*
— *of inventorship (Rule 4.17(iv))*

Published:

— *with international search report (Art. 21(3))*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

METHOD AND SYSTEM FOR TIME-SCALED AUDIOVISUAL FEEDBACK OF SPEECH PRODUCTION EFFORTS

TECHNICAL FIELD

[0001] The present disclosure relates to the field of speech processing, and it more particularly relates to a method and a system for audiovisual feedback of articulatory and prosodic efforts in speech production for speech training.

BACKGROUND

[0002] Children with profound hearing impairment lack auditory feedback during speech production. They consequently experience difficulty in acquiring speech production despite functional articulatory organs. Sign language uses a combination of hand shapes and facial expressions instead of sounds for communication and enables hearing-impaired persons to express themselves. However, most normal-hearing persons do not understand sign language. Speech training aims at helping hearing-impaired persons to produce intelligible speech despite the inability to hear their sounds and thus be able to express themselves to those who do not know sign language. Speech training often uses a mirror for visual feedback on movements of the external articulators (lips, jaw). However, movements of the internal articulators (tongue, vocal-tract shape, glottis, etc.) remain hidden, and no information on voicing, pitch, and level is provided.

[0003] Several speech training systems have been reported using automatic speech recognition, articulatory effort sensors, and a combination. Javkin et al. (H. R. Javkin, E. G. Keate, N. Antonanzas-Barroso, B. A. Hanson, "Synthesis-based speech training system," U.S. Patent No. US 5536171, 1996) described a system to display the model articulatory movements for the target utterance, output the target utterance using a formant synthesizer, provide visual feedback of the learner's articulatory movements using a set of speech transducers (a palatograph to sense the tongue-palate contact pattern, a nasal sensor, a throat sensor, an airflow sensor, and a microphone), and evaluate the similarity of the learner's utterance with the

model utterance. Russell et al. (M. J. Russell, W. S. Robert, L. W. Julie, "Speech training aid," U.S. Patent No. US 5791904, 1998) described a system to assess utterance correctness using stored acoustic models. Komissarchik et al. (J. Komissarchik, E. Komissarchik, "Language independent suprasegmental pronunciation tutoring system and methods," U.S. Patent No. US 6397185, 2002) described a suprasegmental pronunciation tutoring system to display syllabic segmentation along with the waveform, energy, and pitch plots. A system was described by Shpiro et al. (Z. Shpiro, E. Cohen, "Comprehensive spoken language learning system," U.S. Patent Application Publication No. US 2004/0176960, 2004) to provide feedback on errors in pronunciation, stress, and intonation. Handal et al. (A. H. Handal, M. Gary, H. D. Wilson, M. Lessac, "Speech training method with alternative proper pronunciation database," U.S. Patent No. US 6963841, 2005) described a pronunciation assessing system to optionally display the pre-stored articulatory movements as wire-frame drawings. A system was described by Russell (T. Russell, "Method system and software for teaching pronunciation," U.S. Patent Application Publication No. US 20060004567 A1, 2006) using linear prediction to obtain formant trajectories for the learner and teacher's utterances and compares them to provide instructions for pronunciation correction. A pronunciation correction system was described by Rechlis (G. Rechlis, "System and method for correcting speech," U.S. Patent Application Publication No. US 2009/0220926 A1, 2009) using speech recognition and multiple reference recordings of each utterance. An accent correction system was described by Basson et al. (S. H. Basson, D. Kanevsky, E. E. Kelley, B. Ramabhadran, "Method and system for accent correction," U.S. Patent No. US 8175882 B2, 2012) to provide feedback by comparing the speech utterances with the corresponding baseline models. The usefulness of these systems is limited by speech recognition errors and a lack of corrective visual feedback of the articulatory efforts.

[0004] Computer-based speech-training aids providing a dynamic display of acoustic parameters (speech level, voicing, pitch, spectral features, etc.) and articulatory parameters (movements of lips, jaw, tongue, etc.) are reported to be

useful for speech therapy. These aids include 'Speech Viewer', 'Video Voice Training System', 'Dr. Speech', 'Box-of-Tricks', concatenative articulatory video synthesis, 3-D virtual talking head, automatic lipreading recognition system, etc. Several mobile-based apps have been developed in recent years, including 'Voice Analyst', 'Voice Tools', and 'Voice Pitch Analyzer', to display speech level and pitch information for speech therapy. The usefulness of most of these aids is limited as they use language-dependent processing and machine learning to generate visual feedback.

[0005] Despite the development of several speech-training aids, most speech therapists and special education teachers in the schools for hearing-impaired children find it more convenient to use a mirror, improvised gestures, and repeated and extended articulations. Such use indicates a need to co-design the aid with the speech therapists and teachers by interacting with them to understand their difficulties with the available aids and getting their assessment of the features in the new aid. It has been reported by Eriksson et al. (E. Eriksson, O. Engwall, O. Bälter, A. Öster, H. Kjellström, "Design recommendations for a computer-based speech-training system based on end-user interviews," Proceedings of International Conference on Speech and Computer, Patras, Greece, 2005) that optimizing the level of details on display, emphasizing the key articulatory efforts, adapting the feedback to the learner's level, and augmenting with complementary information can improve the usefulness of the aids.

[0006] The speech-training aid users during interactive sessions indicated their preference for a two-panel aid with a variable-rate display. The two-panel display enables visual comparison of the teacher's and the learner's speech production efforts, and the variable-rate feedback permits using a rate suitable for the learning level. It avoids the need for repetitive articulation by the teacher and facilitates remote learning with pre-recorded utterances of a reference speaker. An aid with a two-panel display to display level, pitch, and lateral vocal-tract shape estimated from the speech signal, with a variable-rate animation of the articulatory efforts was reported by Jain et al. (R. Jain, K. S. Nataraj, P. C. Pandey, "Dynamic

display of vocal tract shape for speech training," Proceedings of National Conference on Communications, Guwahati, India, 2016). This system was used and evaluated by a group of speech therapists and teachers. They suggested display layout enhancement, longer signal acquisition duration, and a selection of faces (male/female, young/adult) for vocal-tract shape animation. Further, a playback of the audio signal time-aligned with the variable-rate animation of the articulatory efforts was suggested to improve speech acquisition in children using hearing aids or cochlear implants. Some users indicated a need for feedback on lip movements.

[0007] Thus, there is a need for a speech-training aid with a two-panel display and variable-rate audiovisual feedback of the speech production efforts combining the articulatory and prosodic information obtained by processing the audio signal, time-aligned frontal view of the face, and correspondingly time-scaled audio signal.

OBJECT OF THE INVENTION

[0008] It is the primary object of the present disclosure to provide a system and a method for improving speech production in hearing-impaired children.

[0009] Another object of the present disclosure is to provide a system and a method to help second-language learners improve unfamiliar words' pronunciation.

[0010] Yet another object of the present disclosure is to provide a system and a method for diagnosing voice disorders.

[0011] Yet another object of the present disclosure is to provide a system and a method for aiding rehabilitation therapy for pitch and intonation control.

SUMMARY

[0012] Hearing-impaired children lack auditory feedback and consequently experience difficulty in acquiring speech production. They can benefit from speech-training aids providing visual feedback of key efforts in speech production.

[0013] In the present disclosure, a method and a system are disclosed for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production. The system can be used as a speech-training aid providing variable-rate audiovisual feedback with the information obtained from an audiovisual recording of the speech utterances. The aid has two panels enabling a comparison between the articulatory efforts of the learner and a teacher or a reference speaker. The audio recording is processed to obtain time-varying vocal-tract shape, level, and pitch, and the frontal view of the face is obtained from the video recording. The vocal-tract shape estimation uses inverse filtering based on linear prediction (LP), and the pitch estimation uses glottal epoch detection using the Hilbert envelope for excitation enhancement. The audiovisual feedback comprises a time-scaled animation of the lateral vocal-tract shape, level, and pitch, a time-aligned display of the frontal view of the face, and a correspondingly time-scaled playback of the audio signal. The disclosed system may also help second-language learners improve unfamiliar sounds' pronunciation. It may help diagnose voice disorders and be usable as an aid in rehabilitation therapy for pitch and intonation control.

[0014] In one aspect of the present disclosure, a system is disclosed for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production to compare the efforts of a plurality of persons. The system comprises at least one processor and a plurality of peripherals interfaced with the processor. The processor receives an audiovisual signal comprising each person's speech as the audio signal and the frontal view of each person's face as the video signal from the peripherals and processes the audiovisual signal of each person to provide time-scaled audiovisual feedback. The

processor comprises a signal acquisition module, a speech analysis module, an animation generation module, and a communication and graphical user interface (CGUI) module.

[0015] The signal acquisition module records a speech utterance of the person as the audiovisual signal comprising the audio signal and the video signal, and selects a segment of the audio signal as a selected audio segment and the corresponding segment of the video signal as a selected video segment for speech analysis and animation generation. The speech analysis module receives the selected audio segment, analyzes the selected audio segment to generate an analysis output comprising time-varying level, spectrum, pitch, and lateral vocal-tract shape, and generates time-scaled audio segments with a plurality of settable slowdown factors from the selected audio segment. The animation generation module receives the analysis output, the time-scaled audio segments, and the selected video segment, and generates a time-scaled animation of the time-varying lateral vocal-tract shape and indicators for pitch, level, and place of articulation with the settable slowdown factor from the analysis output and a time-aligned video of the person's face from the selected video segment. The CGUI module interfaces between the signal acquisition module, the speech analysis module, the animation generation module and the peripherals for: acquiring the audiovisual signal and a plurality of inputs from a plurality of persons, communicating with other digital devices or over the internet and storing and reading data, displaying the time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face, and outputting the time-scaled audio signal for a plurality of persons.

[0016] In another aspect of the present disclosure, a method is disclosed for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production to compare the efforts of a plurality of persons. The method is implemented by at least one processor (120) and a plurality of peripherals (110) interfaced with the processor (120). The method comprises the steps of (a) recording a speech utterance of the person as an

audiovisual signal comprising the audio signal and the video signal, (b) selecting a segment of the audio signal as a selected audio segment and the corresponding segment of the video signal as a selected video segment for speech analysis and animation generation, (c) analyzing the selected audio segment to generate an analysis output comprising time-varying level, spectrum, pitch, and lateral vocal-tract shape (d) generating a time-scaled animation of the time-varying lateral vocal-tract shape and a plurality of indicators for pitch, level, and place of articulation with the settable slowdown factor from the analysis output and a time-aligned video of the person's face from the selected video segment, and (e) displaying the time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face and outputting the time-scaled audio signal for a plurality of persons.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The detailed description is described with reference to the accompanying figures.

[0018] Figure 1 illustrates a block diagram of the system in accordance with the present disclosure.

[0019] Figure 2 illustrates a display of the system during signal acquisition in accordance with the present disclosure.

[0020] Figure 3 illustrates a display of the system during speech analysis in accordance with the present disclosure.

[0021] Figure 4 illustrates a display of the system during time-scaled animation in accordance with the present disclosure.

DETAILED DESCRIPTION OF THE INVENTION

[0022] In the present disclosure, a method and a system for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and

prosodic efforts in speech production are disclosed. The system may be used as a speech-training aid for audiovisual feedback with the information obtained from an audiovisual recording of the speech utterances. It has a two-panel display to compare the efforts of the learner with the efforts of a teacher or a reference speaker. It uses language-independent signal processing to provide (i) a time-scaled animation of the lateral vocal-tract shape, level, and pitch, (ii) a time-aligned display of the frontal view of the face, and (iii) a correspondingly time-scaled playback of the audio signal.

[0023] Speech training involves repeated utterances of syllables, words, word clusters, and sentences by the learner and feedback for correct speech production by the teacher. The present invention can assist the learning process and speech training by providing information not visible on the face. The feedback is in the form of slow-motion animation of the vocal-tract shape, pitch, and signal level, obtained by processing the audio signal. The learning process is enhanced by complementary information on lip movements by a time-aligned display of the frontal view of the face so that the learner can relate the various articulatory efforts. For facilitating the learning process, the rate of audiovisual feedback can be altered with the learning level.

[0024] Figure 1 illustrates a block diagram of a system (**100**) for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production to compare the efforts of a plurality of persons in accordance with the present disclosure. The system (**100**) comprises at least one processor (**120**) and a plurality of peripherals (**110**) interfaced with the processor (**120**).

[0025] The peripherals (**110**) may comprise at least one audio input device (**111**), at least one video input device (**112**), at least one audio output device (**113**), at least one video display device (**114**), at least one user interface device (**115**), a communication device (**116**), and a storage device (**117**). The audio input device (**111**), such as a microphone, is configured to acquire a person's speech as an audio

signal. The video input device (**112**), such as a video camera, is configured to acquire the frontal view of the person's face as a video signal. The audio output device (**113**), such as a speaker, is configured to provide auditory feedback. The video display device (**114**) is configured as a plurality of panels to display visual feedback. The user interface device (**115**), such as a keyboard, mouse, or touchpad, is configured to receive a plurality of inputs from a person. The communication device (**116**) is configured to communicate with other digital devices or over the internet. The storage device (**117**), such as a hard disk or flash memory, is configured to store and read data.

[0026] The processor (**120**) receives an audiovisual signal comprising each person's speech as the audio signal and the frontal view of each person's face as the video signal from the peripherals (**110**) and processes each person's audiovisual signal to provide time-scaled audiovisual feedback. The processor (**120**) comprises a communication and graphical user interface (CGUI) module (**121**), a signal acquisition module (**122**), a speech analysis module (**123**), and an animation generation module (**124**). The modules of the processor (**121, 122, 123, 124**) may be hardware modules or may be implemented as software modules.

[0027] The signal acquisition module (**122**) of the processor (**120**) records the speech utterance as an audiovisual signal with the person's speech as an audio signal and the frontal view of the person's face as a video signal. A segment of the audio signal is selected, and the corresponding segment of the video signal is selected for speech analysis and animation generation.

[0028] The speech analysis module (**123**) of the processor (**120**) receives the selected audio segment and generates an analysis output comprising time-varying level, spectrum, pitch, and lateral vocal-tract shape. The speech analysis module (**123**) also generates time-scaled audio segments with a plurality of settable slowdown factors.

[0029] The animation generation module (**124**) of the processor (**120**) receives the analysis output, the time-scaled audio segments, and the video

segment. The animation generation module (**124**) generates an animation of the time-varying lateral vocal-tract shape and a plurality of indicators for pitch, level, and place of articulation, with the settable slowdown factor. Further, the animation generation module (**124**) generates a video of the person's face time-aligned with the animation.

[0030] The CGUI module (**121**) of the processor (**120**) interfaces between the signal acquisition module (**122**), the speech analysis module (**123**), the animation generation module (**124**), and the peripherals (**110**) for (i) acquiring the audiovisual signal and a plurality of inputs from a plurality of persons; (ii) communicating with other digital devices or over the internet and storing and reading data; (iii) displaying the time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face for a plurality of persons; and (iv) outputting the time-scaled audio segment for a plurality of persons. The CGUI module (**121**) interfaces with the video display device (**114**) for simultaneously displaying the time-scaled animation for at least two persons on the video display device (**114**). The CGUI module (**121**) may interface with the user interface device (**115**) for selecting the settable slowdown factor from a set of values.

[0031] In an exemplary embodiment of the present invention, the system (**100**) comprises a plurality of processors and a plurality of peripherals. The processors are configured as the CGUI module (**121**), the signal acquisition module (**122**), the speech analysis module (**123**), and the animation generation module (**124**). In this embodiment, the audio signal is plotted in real time during signal acquisition and superimposed on a plurality of color bands indicating the audio signal volume, for example three colors for low, acceptable, and high volume. The audio signal acquisition is carried out with a sampling frequency, for example set to 10 kHz, adequate for the pitch and vocal-tract shape estimation. A video of the frontal view of the person's face is simultaneously recorded and displayed in real time at a frame rate, for example set to ten frame/s, adequate for visualizing the lip movements during speech production and selected to reduce memory and

computation requirements. The video is displayed in a frame marked with a center and a plurality of vertical and horizontal lines to help the person adjust the face position. The utterances used for speech training are usually shorter than 2 s. In the exemplary embodiment of the present invention, the signal acquisition module (**122**) permits a recording duration, for example 10 s, that allows flexibility for the timing of the utterance and multiple utterances to adjust the level. A segment of the audio signal is selected for analysis by the speech analysis module (**123**). The corresponding video segment is automatically selected. In this embodiment, the signal acquisition module (**122**) may also be used to select an audio segment and a corresponding video segment from a previously recorded audiovisual signal.

[0032] In the same embodiment, the speech analysis module (**123**) analyzes the selected audio segment, generates the analysis output comprising time-varying level, spectrum, pitch, and vocal-tract shape, and generates time-scaled audio segments with a plurality of settable slowdown factors. The pitch estimation uses a glottal epoch detection method using the Hilbert envelope for excitation enhancement. This method is reported to be suitable for speech signals with voice disorders by Dasgupta et al. (H. Dasgupta, P. C. Pandey, K. S. Nataraj, "Epoch detection using Hilbert envelope for glottal excitation enhancement and maximum-sum subarray for epoch marking," IEEE Journal of Selected Topics Signal Processing, 14, pp. 461–471, 2019). The vocal-tract shape estimation uses the inverse filtering method based on linear prediction (LP) to obtain the vocal-tract area function, as reported by Wakita (H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," IEEE Transactions on Audio Electroacoustics, 21, pp. 417–427, 1973).

[0033] In the same embodiment, the analysis by the speech analysis module (**123**) is carried out using a window length of two average pitch periods and a small window shift, for example 5 ms. For improving consistency of the vocal-tract shape estimation, the area values for analysis windows positioned at the minimum windowed energy are linearly interpolated to obtain area values for every window shift, as reported by Nataraj et al. (K. S. Nataraj, Jagbandhu, P. C. Pandey, M. S.

Shah, "Improving the consistency of vocal tract shape estimation," Proceedings of National Conference on Communications, Bangalore, India, 2011). The cubic B-spline interpolation is used to estimate multiple, for example 20, uniformly placed area values along the length of the oral cavity for smoothing the estimated vocal-tract area function. For synchronized audiovisual playback, the selected audio segment is uniformly time-expanded by the set slowdown factor using the synchronous overlap-add with fixed synthesis (SOLAFS) method as reported by Hejna and Musicus (D. Hejna, B. R. Musicus, "The SOLAFS time-scale modification algorithm," Technical Report, University of Cambridge, UK, 1991).

[0034] In the same embodiment, the animation generation module (**124**) processes the analysis output and generates the images with vocal-tract animation and indicators for pitch, level, and place of articulation. A time-aligned video of the frontal view of the person's face is generated to provide simultaneous feedback of articulatory efforts visible on the face. The vocal-tract animation and video of the frontal view of the person's face are displayed using the video display device (**114**) at the set slowdown factor, along with playback of the time-scaled audio signal using the audio output device (**113**).

[0035] In the same embodiment, the estimated area values are used for vocal-tract animation with a two-dimensional mid-sagittal view of the head, as reported by Park et al. (S. H. Park, D. J. Kim, J. H. Lee, T. S. Yoon, "Integrated speech training system for hearing impaired," IEEE Transactions on Rehabilitation Engineering, 2, pp. 189–196, 1994). The estimated oral cavity opening is represented as the area between the fixed upper curve and the moving lower curve. The fixed upper curve comprises the upper lip, upper teeth, and palate, whereas the lower curve comprises the lower lip, lower teeth, and tongue. The place of maximum constriction along the vocal-tract length is indicated as the place of articulation. The vocal-tract area function is estimated with a 5-ms window shift resulting in 200 frame/s for real time animation, for the video of the frontal view of the face available at ten frame/s. The vocal-tract animation, level, and pitch are

time-scaled for the set slowdown factor by down-sampling. The video is time-scaled for the set slowdown factor by frame repetition.

[0036] In the same embodiment, the display of the video display device (**114**) is configured as a plurality of panels, for example two panels to enable a comparison of the articulatory efforts of the learner and the teacher or a reference speaker. Further, the system can simultaneously display the time-scaled animation in the panels. The two-panel display can also be used to validate the animation of the vocal-tract shape using a side-by-side display of the analysis output and animation for the same signal segment. Further, each panel of the display has a vertical menu bar with a plurality of graphical controls for inputs to the signal acquisition, speech analysis, and animation generation modules. The plurality of graphical controls may receive inputs from a person or a plurality of persons. The details on the display are controlled by a plurality of graphical controls on the bottom horizontal and center vertical toolbars. The display of the system of this embodiment during the signal acquisition is shown in Figure 2. Each panel has a plurality of graphical controls for recording (record, start, stop) the audiovisual signal, file opening, segment selection, playback (play/pause, reset) of the selected segment, and saving the selected segment. Further, each panel has a plurality of display sections, for example three sections with the upper section for the audio signal, the middle section for the video of the frontal view of the person's face, and the lower section for the selected audio segment. A segment of the acquired audio signal is selected using movable cursors on the audio signal plot.

[0037] In the same embodiment, the speech analysis module (**123**) displays the time-varying spectrum as a spectrogram. A two-dimensional plot of the time-varying vocal-tract area function is displayed as an 'areagram,' as reported by Pandey et al. (P. C. Pandey, M. S. Shah, "Estimation of place of articulation during stop closures of vowel-consonant-vowel utterances," IEEE Transactions on Audio, Speech, and Language Processing, 17, pp. 277–286, 2009). These two plots are accompanied by plots of the audio signal, level, and pitch. An example display of the speech analysis is shown in Figure 3. The plots of the analysis output in this

display can be used for diagnosing voice disorders. The plot values in either of the panels can be read by placing the cursor at the corresponding position.

[0038] In the same embodiment, the animation generation module (**124**) displays the selected speech segment, frontal view of the face, and vocal-tract shape animation, as shown in Figure 4. A plurality of graphical controls is provided for customizing the display: level and pitch bars, a marker for the place of articulation, the face for animation, and the animation slowdown factor. The animation may use a face of a man, woman, boy, or girl, and a left or right facing. The animation slowdown factor may be selected from a set of values, for example 1, 2, 5, and 10.

[0039] In another embodiment of the present disclosure, the peripherals (**110**) may comprise a plurality of audio input devices, a plurality of video input devices for example video cameras, a plurality of audio output devices, a plurality of video display devices, a plurality of user interface devices.

[0040] The system of the preferred embodiment has been tested for speech analysis output, the module functionalities, the animation correctness for speech utterances from adult male and female speakers, and ease of use of the graphical controls. A processing delay of about 5 s to generate the animation for a 1-s speech segment may be considered acceptable for speech training. The system of the preferred embodiment has a much shorter processing delay. In the system of the preferred embodiment, the settable slowdown factor and the indicators on the display panel can be adapted to the person's learning level. The speech training can be carried out without needing repeated utterances by the teacher.

[0041] The speech-training aid of the present invention can be used for almost any language with the availability of a teacher or a reference speaker. Further, the system does not require specific instruments to be attached to the speaker, and the audiovisual signal acquisition does not interfere with speech production.

[0042] The method and system in this disclosure are primarily aimed at improving the hearing-impaired children's speech production. The system can also be used by second-language learners to improve the unfamiliar sounds' pronunciation. It may also be helpful in the diagnosis of voice disorders and speech rehabilitation therapy for pitch and intonation control.

[0043] The above description and the accompanying drawings are intended to describe an exemplary embodiment of the present invention in sufficient detail to enable those skilled in the art to practice the invention. This description should not be interpreted as limiting the scope of the invention because various embodiments with changes in form and detail are possible without departing from the spirit and scope of the disclosure. For example, the processing modules (**121**, **122**, **123**, **124**) described in the disclosure may be partitioned and/or combined in many ways. The processor (**120**) may be a general-purpose processor with a non-transient memory or other storage medium, a digital signal processor, an embedded microcontroller, an FPGA (field programmable gate array), an ASIC (application-specific integrated circuit), or a combination of a plurality of such processors, with one or more modules implemented using one processor. The peripheral devices (**111**, **112**, **113**, **114**, **115**, **116**, **117**) may be substituted by devices combining the functionalities of a plurality of these devices. For example, the video display device (**114**) and the user interface device (**115**) may be substituted by a touchscreen. The menu bars and graphical controls on the display panels may be arranged in many ways for the same functionality.

We claim:

1. A system (100) for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production to compare the efforts of a plurality of persons, the system (100) comprising at least one processor (120) and a plurality of peripherals (110) interfaced with the processor (120), wherein the peripherals (110) comprise:

(a) at least one audio input device (111) configured to acquire a person's speech as an audio signal;

(b) at least one video input device (112) configured to acquire the frontal view of the person's face as a video signal;

(c) at least one audio output device (113) configured to provide auditory feedback;

(d) at least one video display device (114) configured as a plurality of panels to display visual feedback;

(e) at least one user interface device (115) configured to receive a plurality of inputs from the person;

(f) a communication device (116) configured to communicate with other digital devices or over the internet; and

(g) a storage device (117) configured to store and read data;

and the processor (120) receives an audiovisual signal comprising each person's speech as the audio signal and the frontal view of each person's face as the video signal from the peripherals (110) and processes the audiovisual signal of each person to provide time-scaled audiovisual feedback, wherein the processor (120) comprises:

(a) a signal acquisition module (122) configured to:

(i) record a speech utterance of the person as the audiovisual signal comprising the audio signal and the video signal, and

(ii) select a segment of the audio signal as a selected audio segment and the corresponding segment of the video signal as a selected video segment for speech analysis and animation generation;

(b) a speech analysis module (123) configured to:

(i) receive the selected audio segment,

(ii) analyze the selected audio segment to generate an analysis output comprising time-varying level, spectrum, pitch, and lateral vocal-tract shape, and

(iii) generate time-scaled audio segments with a plurality of settable slowdown factors from the selected audio segment;

(c) an animation generation module (124) configured to:

(i) receive the analysis output, the time-scaled audio segments, and the selected video segment, and

(ii) generate a time-scaled animation of the time-varying lateral vocal-tract shape and indicators for pitch, level, and place of articulation with the settable slowdown factor from the analysis output and a time-aligned video of the person's face from the selected video segment; and

(d) a communication and graphical user interface (CGUI) module (121) configured to interface between the signal acquisition module (122), the speech analysis module (123), the animation generation module (124), and the peripherals (110) for:

(i) acquiring the audiovisual signal and a plurality of inputs from a plurality of persons;

(ii) communicating with other digital devices or over the internet and storing and reading data;

(iii) displaying the time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face for a plurality of persons; and

(iv) outputting the time-scaled audio signal for a plurality of persons.

2. The system as claimed in claim 1, wherein each panel of the video display device (114) provides a plurality of graphical controls for receiving the inputs from a person.

3. The system as claimed in claim 1, wherein each panel of the video display device (114) is configured to display the audiovisual signal in real time during signal acquisition with:

(a) the audio signal plot superimposed on a plurality of color bands to indicate the audio signal volume; and

(b) the frontal view of the person's face displayed in a frame marked with a center and a plurality of vertical and horizontal lines to adjust the face position.

4. The system as claimed in claim 1, wherein the audio signal acquisition is carried out with a settable sampling frequency for estimating the pitch and vocal-tract shape, and the video of the face is acquired at a settable frame rate for visualizing the lip movements during speech production.

5. The system as claimed in claim 1, wherein the signal acquisition module (122) is configured to select an audio segment and the corresponding video segment optionally from a previously recorded audiovisual signal.

6. The system as claimed in claim 1, wherein the CGUI module (121) is configured to simultaneously display the time-scaled animation for at least two persons on the video display device (114).

7. The system as claimed in claim 1, wherein the CGUI module (121) is configured to select the settable slowdown factor from a set of values.

8. The system as claimed in claim 2, wherein the graphical controls are configured to:

(a) record the audiovisual signal;

(b) select a segment of the audiovisual signal;

(c) playback the selected segment of the audiovisual signal;

(d) save the selected segment of the audiovisual signal; and

(e) customize the display with level and pitch bars, a marker for the place of articulation, a face for animation, a direction of the face, and the animation slowdown factor.

9. A method for automatically generating a variable-rate time-scaled audiovisual feedback of articulatory and prosodic efforts in speech production to compare the efforts of a plurality of persons, wherein the method is implemented by at least one processor (120) and a plurality of peripherals (110) interfaced with the processor (120) and the method comprises the steps of:

(a) recording a speech utterance of the person as an audiovisual signal comprising the audio signal and the video signal;

(b) selecting a segment of the audio signal as a selected audio segment and the corresponding segment of the video signal as a selected video segment for speech analysis and animation generation;

(c) analyzing the selected audio segment to generate an analysis output comprising time-varying level, spectrum, pitch, and lateral vocal-tract shape;

(d) generating time-scaled audio segments with a plurality of settable slowdown factors from the selected audio segment;

(e) generating a time-scaled animation of the time-varying lateral vocal-tract shape and a plurality of indicators for pitch, level, and place of articulation with the settable slowdown factor from the analysis output and a time-aligned video of the person's face from the selected video segment; and

(f) displaying the time-scaled vocal-tract animation, indicators for pitch, level, and place of articulation, and the time-aligned video of the person's face and outputting the time-scaled audio signal for a plurality of persons.

10. The method as claimed in claim 9, wherein the audiovisual signal and a plurality of inputs are acquired from a plurality of persons.

11. The method as claimed in claim 9, wherein the audiovisual signal is displayed in real time with:

(a) the audio signal plot superimposed on a plurality of color bands to indicate the audio signal volume; and

(b) the frontal view of the person's face displayed in a frame marked with a center and a plurality of vertical and horizontal lines to adjust the face position.

12. The method as claimed in claim 9, wherein the time-scaled animation for at least two persons are simultaneously displayed.

13. The method as claimed in claim 9, wherein an audio segment and a corresponding video segment are selected optionally from a previously recorded audiovisual signal.

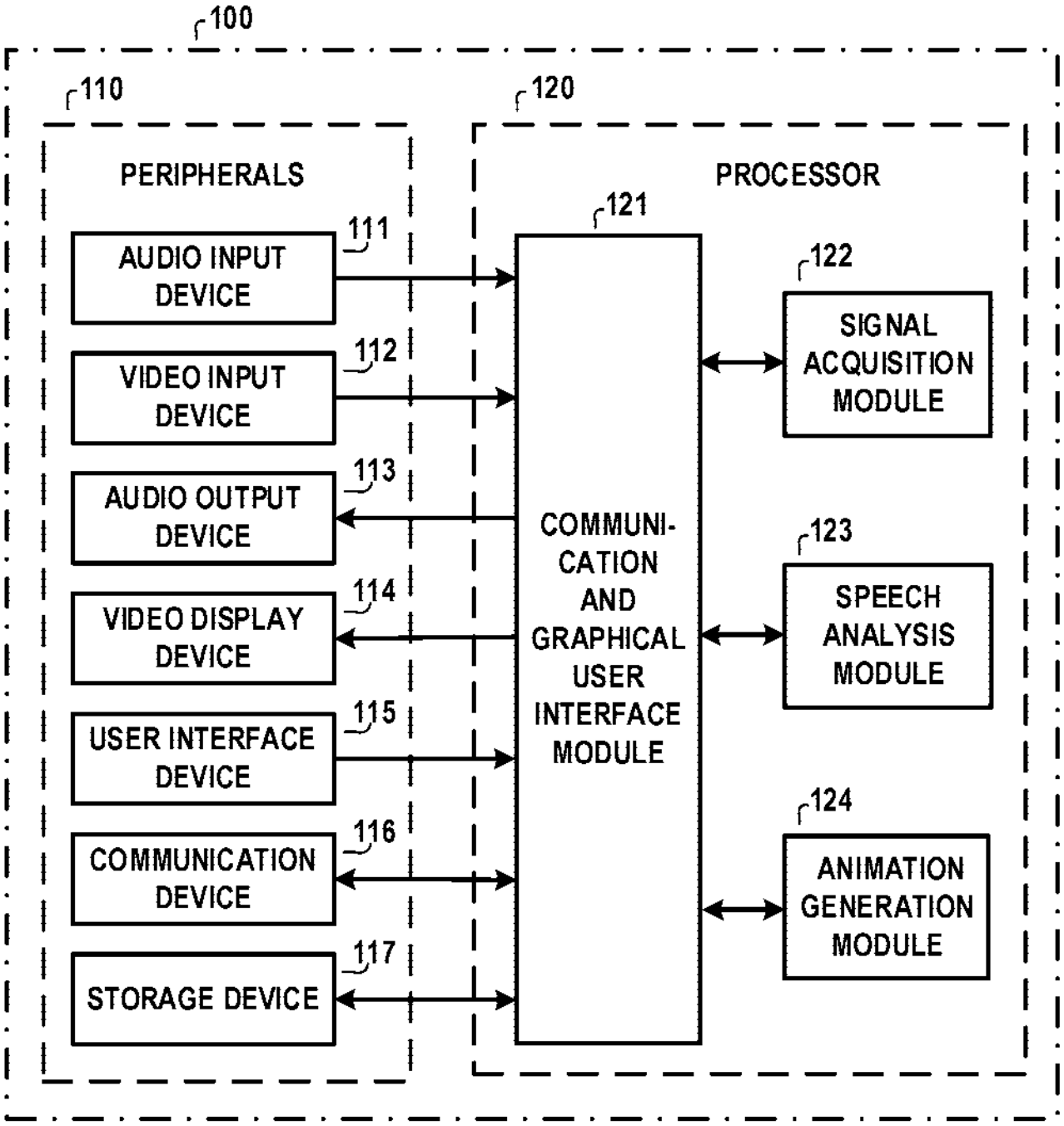
100
110
PERIPHERALS
AUDIO INPUT DEVICE
111
VIDEO INPUT DEVICE
112
AUDIO OUTPUT DEVICE
113
VIDEO DISPLAY DEVICE
114
USER INTERFACE DEVICE
115
COMMUNICATION DEVICE
116
STORAGE DEVICE
117
120
PROCESSOR
121
COMMUNI-CATION AND GRAPHICAL USER INTERFACE MODULE
122
SIGNAL ACQUISITION MODULE
123
SPEECH ANALYSIS MODULE
124
ANIMATION GENERATION MODULE

Figure 1

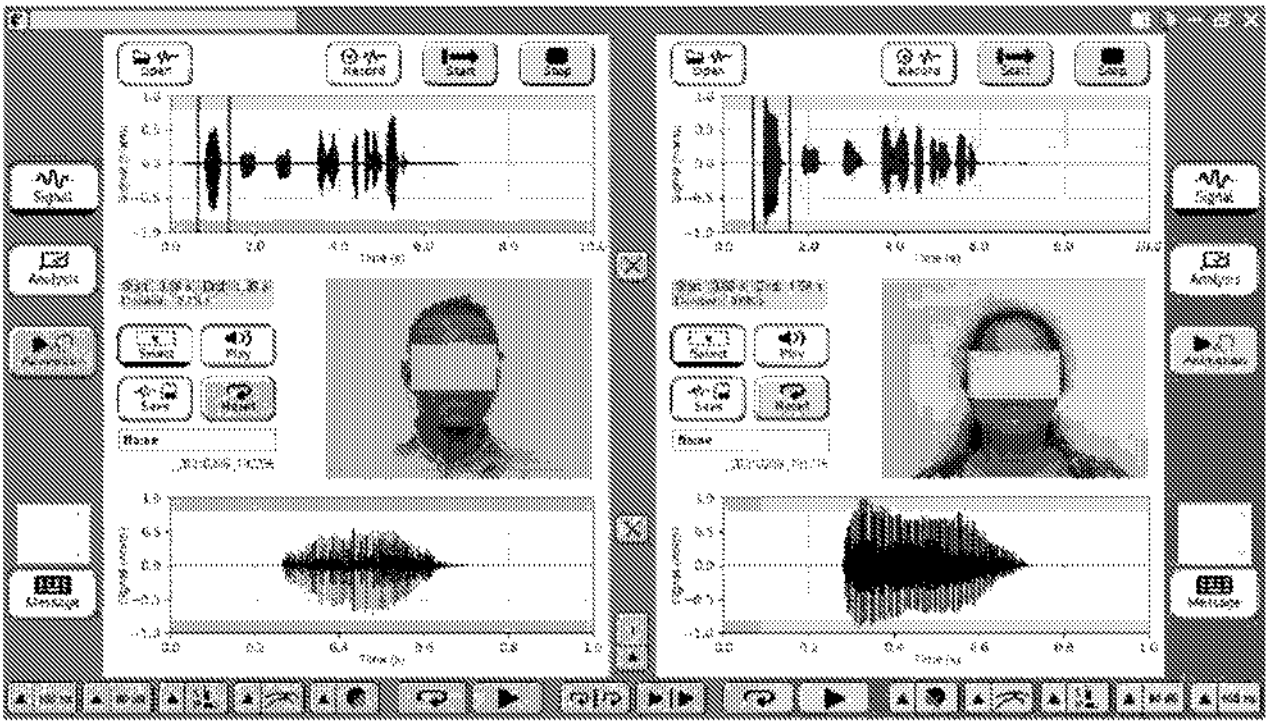

Figure 2

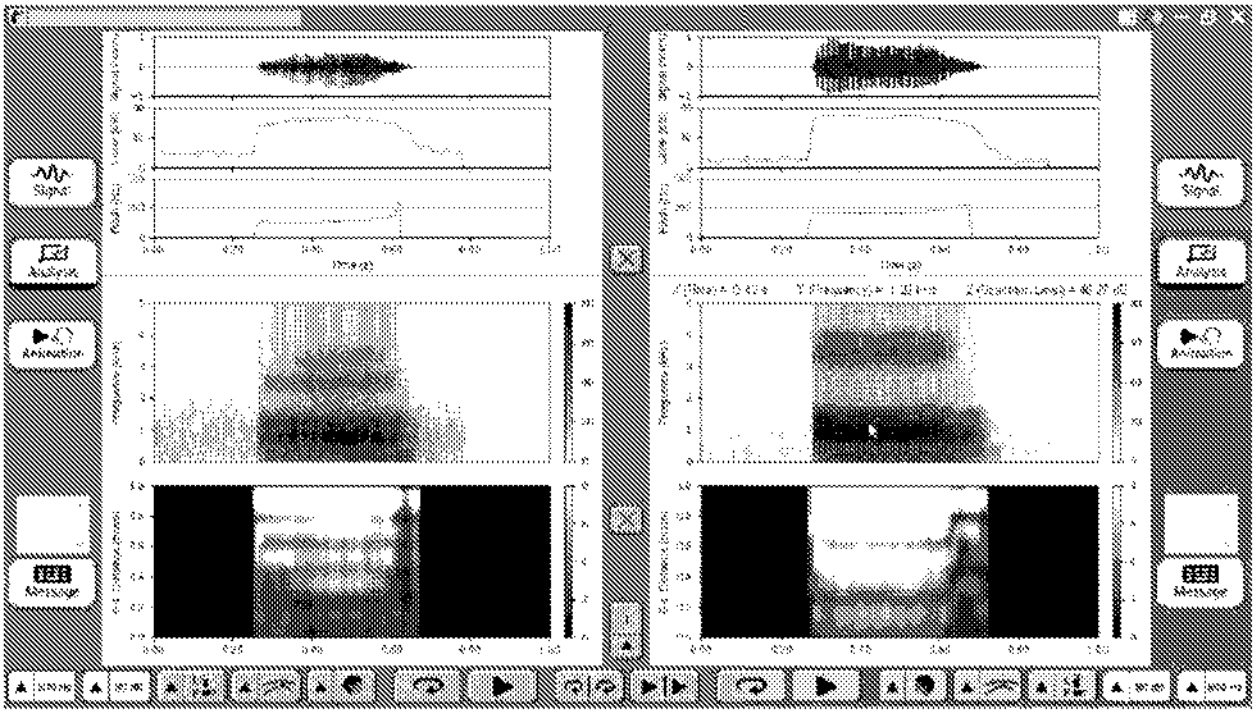

Figure 3

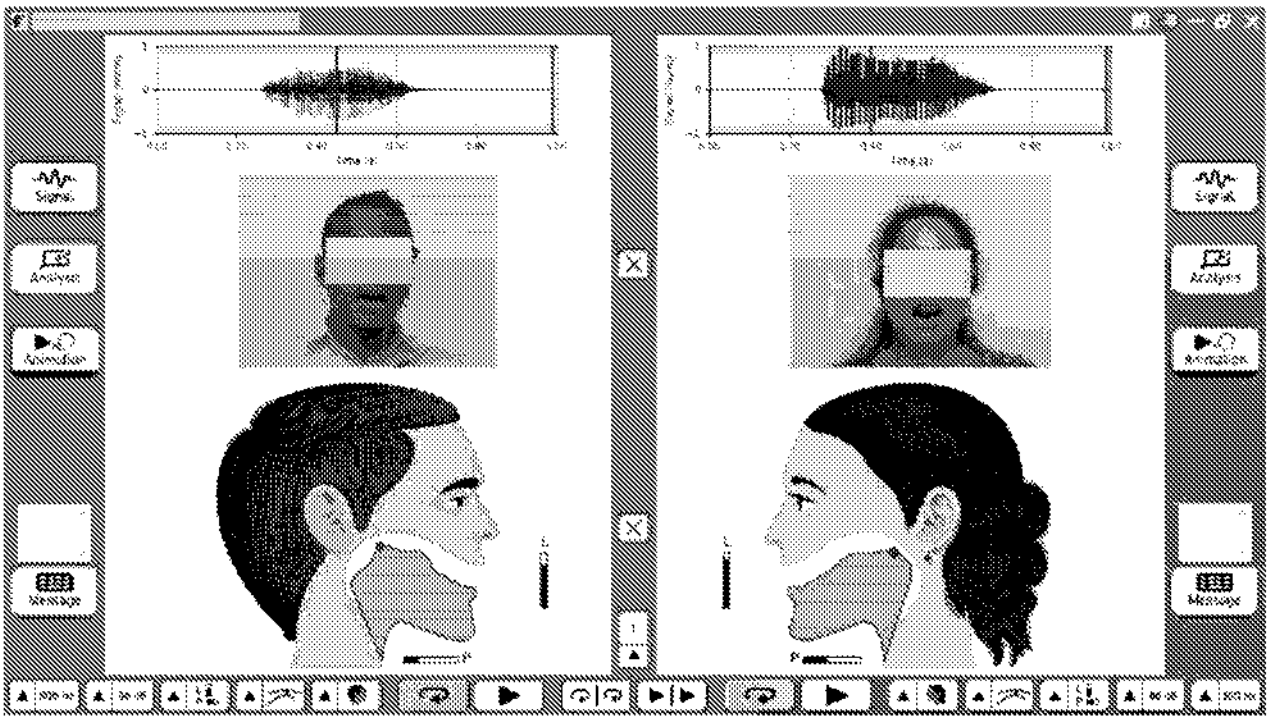

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/IN22/50658

A. CLASSIFICATION OF SUBJECT MATTER

IPC - INV. G06V 40/16; G09B 19/04; G10L 21/0356; G10L 21/12; G10L 21/18; G10L 25/72; G10L 25/75; G09B 5/04 (2022.01)
ADD. G09B 19/06; G10L 21/0332; G10L 21/043; G10L 21/055 (2022.01)

CPC - INV. G06V 40/166; G09B 19/04; G10L 21/0356; G10L 21/12; G10L 21/18; G10L 25/72; G10L 25/75; G06V 40/171; G09B 5/04; G10L 21/0332; G10L 21/043; G10L 21/055
ADD. A61B 5/4803; G09B 19/06; G10L 2021/0575; G10L 2021/065

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic database consulted during the international search (name of database and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016/0321953 A1 (KANG, J) 03 November 2016; Fig. 1, 40, 56, 58-60, 64; para [0012]-[0022], [0049]-[0061], [0110]-[0127], [0131]-[0135], [0143]-[0162], [0170]-[0189]	1-13
A	US 2002/0087322 A1 (FLETCHER, S) 04 July 2002; entire document	1-13
A	US 2013/0035940 A1 (WAN, M ET AL.) 07 February 2013; entire document	1-13

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"D" document cited by the applicant in the international application
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
12 October 2022 (12.10.2022)	DEC 05 2022
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas Telephone No. PCT Helpdesk: 571-272-4300