CLAIMS

1. An audio processing device, in particular a mobile electronic user device, an audio processing device integrated into a wearable hearing system or a server, for improving the flow of words in fluency disorders, in particular in the case of stuttering, wherein the audio processing device is configured for:

receiving input audio information from an audio sensor device, which comprises at least one verbal utterance in a natural voice of a user;

carrying out a voice conversion for generating output audio information in an ego-dystonic target voice, wherein the ego-dystonic target voice is a voice, which is identified by the user as a foreign voice of another speaker in a sensory or neural manner by means of a neural mechanism of the auditory cortex for identifying the voice of the user, wherein carrying out the voice conversion comprises:

calculating a quantifiable self-dissimilarity value for each of multiple target voices using a Euclidean distance method where the target voices and the user's voice are each represented as points in a multidimensional acoustic feature space;

performing an algorithm to determine maximal dissimilarity from the user's voice, wherein maximal dissimilarity is quantitatively determined by one or more of the following criteria:

- a dissimilarity value that surpasses a predetermined threshold value as quantified by the Euclidean distance metric;
- a percentile ranking where the target voice is positioned within the top 30% in terms of distance from the user's voice, as calculated in a voice similarity matrix;
- a dissimilarity value that exceeds a specific number of standard deviations above the mean value in the dataset;

based on the determined maximal dissimilarity from the user's natural voice, manipulating formant frequencies and/or spectral features and/or timbre while maintaining the natural fundamental frequency f0 of the voice of the user, thereby converting the at least one verbal utterance as if the same speech content was produced by a different speaker; and

prompting a reproduction, in particular a binaural reproduction, of the voice-converted output audio information to the user as acoustic feedback to the speaking of the user;

wherein the audio processing device is further configured for continuously changing a voice identity of the target voice to prevent adaptation by the user.

2. The audio processing device for the use according to claim 1,

wherein the ego-dystonic target voice is a voice, which maintains the pitch of the natural voice of the user; and/or

wherein the ego-dystonic target voice is a voice, which has a lifelike or at least approximately lifelike voice naturalness; and/or

wherein the ego-dystonic target voice is a voice, which maintains or approximately maintains the natural quality of a human voice; and/or

wherein the ego-dystonic target voice is a voice, which is not solely based on a change of the pitch and/or a modification by means of frequency filtering.

3. The audio processing device for the use according to claim 1 or 2, wherein the audio processing device for the voice conversion is at least partially configured based on a machine learning model;

wherein the machine learning model comprises a deep neural network (DNN), a recurrent neural network (RNN), a generative adversarial network (GAN) and/or a sequence-to-sequence mapping network (S2S); and/or

wherein the machine learning model is configured to carry out one or several of the following operations:

reproducing individual natural and/or synthetic speaker voices, which are used for the machine learning;

generating new speaker voices, which are not used for the machine learning; and/or

wherein the voice conversion, or at least parts thereof, is carried out in a language-dependent manner (intra-lingually) or cross-lingually; and/or

wherein the voice conversion, or at least parts thereof, is carried out in a gender-dependent manner (intra-gendered) or in a gender-independent manner (cross-gendered).

4. The audio processing device for the use according to any one of the preceding claims, wherein the ego-dystonic target voice is a voice, which deviates in at least one of the following features from the natural voice of the user: stretching, shortening, widening, constriction of the physiological vocal tract of the user.

5. The audio processing device for the use according to any one of the preceding claims, wherein the ego-dystonic target voice comprises an anti-voice, which maximally deviates from the natural voice of the user in at least one speaker-dependent, non-lingual voice feature;

wherein the at least one voice feature comprises:

one or several speaker-dependent spectral properties, which depend directly on the configuration of the vocal tract, such as, for example, “Mel-frequency cepstral coefficients (MFCCs)“, “linear prediction cepstral coefficients (LPCCs)“ and/or “perceptual linear prediction coefficients, and/or

one or several speaker-dependent prosodic properties, such as, for example, “instantaneous energy“, “intonation“, “speech rate“, and/or “unit durations“; and/or

one or several speaker-dependent features of the way of speaking, in particular of the linguistic dialect.

6. The audio processing device according to any one of the preceding claims, wherein:

- a percentile ranking where the anti-voice is positioned within the top 20%, and more preferably within the top 10%, in terms of distance from the user’s voice, as calculated in a voice similarity matrix;
- a dissimilarity value that exceeds a specific number of standard deviations above the mean value in the dataset, namely one standard deviation, and more preferably two standard deviations;
- a comparative measure where the dissimilarity of the anti-voice is classified as falling within the highest 30% or the farthest 30% of all target-voices from the user’s voice within a voice similarity matrix;
- an absolute cut-off point, established based on empirical evidence or expert consensus, beyond which the dissimilarity score of any voice is considered maximal for the purpose of human sensory-based voice recognition.

7. The audio processing device according to any one of the preceding claims, wherein the audio processing device is further configured to perform an algorithm to determine maximal dissimilarity from the user’s natural voice in at least one acoustic parameter critical for the human neural or sensorial recognition of a voice identity, and wherein maximal dissimilarity is determined by one or more of the following criteria:

- the respective parameter exhibits maximal dissimilarity if its measurement surpasses a pre-established threshold in the Euclidean distance metric, signifying a substantial deviation from the respective parameter in the user's natural voice;

- the respective parameter exhibits maximal dissimilarity if ranked within the top 30%, preferably within the top 20%, and most preferably within the top 10%, in terms of its distance from the corresponding parameter of the user's voice, as assessed in a parameter-specific similarity matrix;

- the respective parameter exhibits maximal dissimilarity when its value is several standard deviations above the dataset's mean for that specific parameter, preferably one standard deviation, and more preferably two standard deviations;

- a comparative measure is utilized where maximal dissimilarity in the respective parameter is classified when it is ranked within the highest 30% or as one of the farthest 30% measurements from the user's corresponding voice parameter in a parameter-specific similarity matrix;

- an absolute cut-off point, established based on empirical evidence or expert consensus, beyond which the dissimilarity score of the respective parameter is considered maximal for the purpose of human sensory-based voice recognition;

and wherein the conversion of the at least one acoustic parameter is carried out based on the determined maximal dissimilarity, by using a voice conversion model.

8. The audio processing device for the use according to any one of the claims 5 to 7, wherein the audio processing device is further configured for:

determining at least one user-specific vocal feature, such as gender, age, features of the vocal tract and/or linguistic dialect in a setup phase, for example on the basis of at least one speech sample;

converting the at least one feature by using a voice conversion model, which is in particular based on machine learning, when carrying out the voice conversion;

wherein the conversion comprises at least one of:

converting a male to a female voice and/or vice versa; and/or

converting an old to a young voice and/or vice versa; and/or

converting a stretched to a shortened vocal tract and/or vice versa; and/or

converting a wide to a narrow vocal tract and/or vice versa; and/or

converting a linguistic dialect, for example a Northern English dialect to a Southern English dialect.

9. The audio processing device for the use according to any one of the preceding claims, wherein the audio processing device is further configured for: carrying out a voice anonymization or voice pseudo-anonymization for concealing the voice identity of the user in the output audio information.

10. The audio processing device for the use according to any one of the preceding claims, wherein the audio processing device for the voice conversion for generating output audio information is further configured for:

capturing information relating to the head position, location and/or movements of the user, in particular by means of a wearable hearing system used by the user; and

using the captured information in order to add spatial audio references during the step of the reproduction of the voice-converted output audio information, which audio references are created by algorithms of the 3D position audio for virtually placing sound sources at any point in the three-dimensional space, such as the “head-related transfer function“, which give an auditory impression as if the target voice originates from a predetermined ego-dystonic position within a three-dimensional acoustic space, for example, behind, above, in front of or below the user.

11. The audio processing device for the use according to any one of the preceding claims, wherein the continuous change takes place so that a hearing impression of the target voice remains novel for the user; and/or

wherein the continuous change takes place according to a constant change rate G, at which the target voice changes step-by-step; and/or

wherein the change rate G corresponds to a speed, at which a first voice identity transitions completely into a second, perceivably different voice identity, wherein the change rate G is expressed in percent per second; and/or

wherein the change rate G is determined so that the changing takes place inconspicuously and/or below a perception threshold for acoustic changes.

12. The audio processing device for the use according to any one of the preceding claims, wherein the audio processing device is further configured for:

in response to detecting a speech activity of the user, dividing sections of the input audio information into sections with speech and sections without speech, wherein the

execution of the voice conversion to generate the output audio information is only based on the sections with speech;

wherein, optionally, the division is performed by means of a machine learning model; and

wherein, optionally, data of a structure-borne sound-related audio sensor system is observed and classified beforehand in order to differentiate between vocal activity and non-vocal activity of the user, wherein only the part of the input audio information identified as vocal activity is transferred to the speech activity recognition.

13. The audio processing device for the use according to any one of the preceding claims, wherein the audio processing device is further configured for:

adding a digital water mark, which is imperceptible for the user, to the output audio information in order to make the target voice identifiable as artificially changed voice, in particular for systems and methods of voice identification.

14. The audio processing device for the use according to any one of the preceding claims, wherein the audio processing device is configured to continuously improve the ego-dystonic target voice based on input audio information characterizing the user's speech fluency.

15. The audio processing device for the use according to claim 14, wherein the ego-dystonic target voice is altered based on a quantification of the improvement in the user's speech fluency when exposed to various ego-dystonic target voices or anti-voices.

16. The audio processing device for the use according to claim 14 or 15, wherein the audio processing device is configured to apply a machine learning model, in particular a machine learning-based, self-adaptive model, wherein the model is configured to:

- dynamically alter one or more audio parameters of the target voice, including but not limited to formant frequencies (timbre) and harmonics, in response to indicators of speech fluency; and/or

- continuously refine the one or more audio parameter adjustments based on successful speech outcomes, such as a quantified reduction in the frequency of stuttering events for the specific user or user group; and/or

- utilize an iterative process for customizing the voice conversion for each individual, thereby optimizing speech fluency through personalized auditory manipulation.

17. A wearable hearing device for improving the flow of words in fluency disorders, in particular in the case of stuttering, comprising:

an audio sensor device for capturing input audio information, which comprises at least one verbal utterance in a natural voice of a user;

means for transferring the input audio information to an audio processing device and for receiving voice-converted output audio information from the audio processing device in an ego-dystonic target voice, in that the at least one verbal utterance has been converted as if the same speech content was produced by a different speaker by manipulating formant frequencies and/or spectral features and/or timbre while maintaining the natural fundamental frequency f0 of the voice of the user;

wherein the ego-dystonic target voice is a voice, which is identified by the user as a foreign voice in a sensory or neural manner by means of a neural mechanism of the auditory cortex for identifying the voice of the user; and

an audio output device for reproducing, in particular binaural reproducing, the voice-converted output audio information to the user as acoustic feedback to the speaking of the user; and

wherein the audio processing device is an audio processing device according to any one of the preceding claims 1-16.

18. A voice conversion system for improving the flow of words in fluency disorders, in particular in the case of stuttering, comprising:

an audio processing device according to any one of the preceding claims 1-16; and

a wearable hearing device according to claim 17.

19. A computer-implemented voice conversion method for improving the flow of speech in the case of fluency disorders, in particular in the case of stuttering, wherein the method is carried out by means of an audio processing device , in particular by means of a mobile electronic user device, by means of an audio processing device integrated into a wearable hearing system or by means of a server, wherein the method comprises at least the following steps:

receiving input audio information from an audio sensor device , which comprises at least one verbal utterance in a natural voice of a user;

carrying out a voice conversion for generating output audio information in an ego-dystonic target voice, wherein the ego-dystonic target voice is a voice, which is identified by the user as foreign voice of another speaker in a sensory or neural manner by means of a

neural mechanism of the auditory cortex for identifying the voice of the user, wherein carrying out the voice conversion comprises:

calculating a quantifiable self-dissimilarity value for each of multiple target voices using a Euclidean distance method where the target voices and the user's voice are each represented as points in a multidimensional acoustic feature space;

performing an algorithm to determine maximal dissimilarity from the user's voice, wherein maximal dissimilarity is quantitatively determined by one or more of the following criteria:

- a dissimilarity value that surpasses a predetermined threshold value as quantified by the Euclidean distance metric;
- a percentile ranking where the target voice is positioned within the top 30% in terms of distance from the user's voice, as calculated in a voice similarity matrix;
- a dissimilarity value that exceeds a specific number of standard deviations above the mean value in the dataset;

based on the determined maximal dissimilarity from the user's natural voice, manipulating formant frequencies and/or spectral features and/or timbre while maintaining the natural fundamental frequency f0 of the voice of the user, thereby converting the at least one verbal utterance as if the same speech content was produced by a different speaker; and

prompting a reproduction, in particular a binaural reproduction, of the voice-converted output audio information to the user as acoustic feedback to the speaking of the user;

wherein the method further comprises continuously changing a voice identity of the target voice to prevent adaptation by the user;

wherein, optionally, the method further comprises one of the several steps, which correspond to the function of the audio processing device according to any one of the preceding claims 2-16.

20. A computer-implemented training method for improving the flow of words and/or speech of a user, wherein the method is carried out by means of an audio processing device , in particular by means of a mobile electronic user device, by means of an audio processing device integrated into a wearable hearing system or by means of a server, wherein the method comprises at least the following steps:

receiving input audio information from an audio sensor device , which comprises at least one verbal utterance in a natural voice of the user;

carrying out a voice conversion for generating output audio information in an ego-dystonic target voice, wherein the ego-dystonic target voice is a voice, which is identified by the user as foreign voice of another speaker in a sensory or neural manner by means of a neural mechanism of the auditory cortex for identifying a voice of the user, wherein carrying out the voice conversion comprises:

calculating a quantifiable self-dissimilarity value for each of multiple target voices using a Euclidean distance method where the target voices and the user's voice are each represented as points in a multidimensional acoustic feature space;

performing an algorithm to determine maximal dissimilarity from the user's voice, wherein maximal dissimilarity is quantitatively determined by one or more of the following criteria:

- a dissimilarity value that surpasses a predetermined threshold value as quantified by the Euclidean distance metric;
- a percentile ranking where the target voice is positioned within the top 30% in terms of distance from the user's voice, as calculated in a voice similarity matrix;
- a dissimilarity value that exceeds a specific number of standard deviations above the mean value in the dataset;

based on the determined maximal dissimilarity from the user's natural voice, manipulating formant frequencies and/or spectral features and/or timbre while maintaining the natural fundamental frequency f0 of the voice of the user, thereby converting the at least one verbal utterance as if the same speech content was produced by a different speaker; and

prompting a reproduction, in particular a binaural reproduction, of the voice-converted output audio information to the subject as acoustic feedback to the speaking of the user;

wherein the method further comprises continuously changing a voice identity of the target voice to prevent adaptation by the user;

wherein, optionally, the training method further comprises one of the several steps, which correspond to the function of the audio processing device according to any one of the preceding claims 2-16.

21. A computer program or a computer-readable memory medium, on which the computer program is stored, wherein the computer program comprises commands, which, when executing the program by a computer, prompt the latter to execute the method according to claim 19 or 20.