(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 3 113 180 B1

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
**22.01.2020 Bulletin 2020/04**

(51) Int Cl.:
***G10L 19/005*** *(2013.01)*

(21) Application number: **15306085.0**

(22) Date of filing: **02.07.2015**

(54) **METHOD FOR PERFORMING AUDIO INPAINTING ON A SPEECH SIGNAL AND APPARATUS FOR PERFORMING AUDIO INPAINTING ON A SPEECH SIGNAL**

VERFAHREN ZUR DURCHFÜHRUNG EINER AUDIO-EINBLENDUNG IN EIN SPRACHSIGNAL UND VORRICHTUNG ZUR DURCHFÜHRUNG EINER AUDIO-EINBLENDUNG IN EIN SPRACHSIGNAL

PROCÉDÉ ET APPAREIL PERMETTANT D'EFFECTUER DES RETOUCHES AUDIO SUR UN SIGNAL VOCAL

(84) Designated Contracting States:
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR**

(43) Date of publication of application:
**04.01.2017 Bulletin 2017/01**

(73) Proprietor: **InterDigital CE Patent Holdings**
**75017 Paris (FR)**

(72) Inventors:
- **Prablanc, Pierre**
  **69300 Caluire (FR)**
- **Duong, Quang Khanh Ngoc**
  **35700 Rennes (FR)**
- **Ozerov, Alexey**
  **35000 Rennes (FR)**
- **Perez, Patrick**
  **92130 ISSY LES MOULINEAUX (FR)**

(74) Representative: **Tarquis-Guillou, Anne et al**
**InterDigital CE Patent Holdings**
**20, rue Rouget de Lisle**
**92130 Issy-les-Moulineaux (FR)**

(56) References cited:
**US-A1- 2011 165 912** **US-A1- 2015 023 345**

- **AMIR ADLER ET AL: "Audio Inpainting", IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, IEEE SERVICE CENTER, NEW YORK, NY, USA, vol. 20, no. 3, 1 March 2012 (2012-03-01), pages 922-932, XP011397627, ISSN: 1558-7916, DOI: 10.1109/TASL.2011.2168211**



EP 3 113 180 B1

## Description

### Field of the invention

**[0001]** The present principles relate to a method for performing audio inpainting on a speech signal, and an apparatus for performing audio inpainting on a speech signal.

### Background

**[0002]** Audio inpainting is the problem of recovering audio samples which are missing or distorted due to, e.g., lost IP packets during a Voice over IP (VoIP) transmission or any other kind of deterioration. Audio inpainting algorithms have various applications ranging from IP packets loss recovery, especially in VoIP or mobile phone, or voice censorship cancelling to various types of damaged audio repairs, including declipping and declicking. Moreover, inpainting might be used for speech modification, e.g., to replace a word of a sequence of words in a speech sequence by some other words. While signal completion has been thoroughly investigated for image and video inpainting, it is much less the case in the context of audio data in general and speech in particular.

**[0003]** Adler et al. [1] introduced an audio inpainting algorithm for the specific purpose of audio declipping, i.e., intended to recover missing audio samples in the time domain that were clipped due to, e.g., limited range of the acquisition device. Also techniques for filling missing segments in the time-frequency domain have been developed [2,3]. However, these methods are not suitable in case of large spectral holes, especially when all frequency bins are missing in certain time frames. Drori et al. [4] proposed another approach to audio inpainting in the spectral domain, relying on exemplar spectral patches taken from the known part of the spectrogram. Bahat et al. [7] proposed a method for filling moderate gaps, e.g. corresponding to the loss of several successive IP packets, and especially in the case of speech signals. These approaches are based on self-similarity of some speech features within the signal and thus perform poorly if the missing part is actually very different from the rest. The known approaches, including the speech-specific method in [7], are unable to cope with situations where quite large temporal gaps are missing in a speech signal. For example, one such gap can cover one entire word or a sequence of words. Indeed, methods based on audio patch similarity or speech feature similarities are unable to recreate entire missing words.

### Summary of the Invention

**[0004]** A novel method to fill gaps in speech data while preserving speech meaning and voice characteristics is disclosed in claim 1.

**[0005]** It has been found that, if a gap occurs in a speech signal, it is very helpful to use any kind of information in order to fill the gap, and that it is possible to fill the gap by using a text transcript of the corresponding utterance. The disclosed speech audio inpainting technique plausibly recovers speech parts that are lost due to, e.g., specific audio editing or lossy transmission with the help of synthetic speech generated from the text transcript of the missing part. The synthesized speech is modified based on conventional voice conversion (e.g., as in [5]) to fit with the original speaker's voice.

**[0006]** A text transcript of the missing speech part is generated or given, e.g., it can be provided by a user, infered by natural language processing techniques based on the known phrases before and/or after the gap, or available from any other source. The text transcript of the missing speech part is used to complete an obfuscated speech signal. It allows leveraging recent progress of text-to-speech (TTS) synthesizers at generating very natural and high quality speech data.

**[0007]** In principle, a method for speech inpainting comprises synthesizing speech for a gap that occurs in a speech signal using a transcript of the speech signal, converting the synthesized speech by voice conversion according to the original speech signal, and blending the synthesized converted speech into the original speech signal to fill the gap.

**[0008]** An apparatus for performing speech inpainting on a speech signal is disclosed in claim 11. The apparatus comprises a speech analyzer that is adapted for detecting a gap in the speech signal, a speech synthesizer that is adapted for performing automatic speech synthesis from text transcript at least for the gap, a voice converter that is adapted for performing voice conversion to adapt the synthesized speech to an original speaker's voice, and a mixer that is adapted for blending of the converted synthesized speech into the original speech audio track. In one embodiment of the mixer, temporal and/or phase mismatches are removed.

**[0009]** Voice conversion is a process that transforms the speech signal from the voice of one person, which is called source speaker, as if it would have been uttered by another person, which is called target speaker. In a usual voice conversion workflow, two steps have to be considered: a learning step and a conversion step. During the learning step, a mapping function is learned to map voice parameters of a source speaker to voice parameters of a target speaker. To model differences between the two speakers, some training data from both speakers are needed. For conversion within the same language, it is more conventional to use parallel training data, which is a set of sentences uttered by both source and target speakers. In the present case, the target speaker is the one whose data are missing whereas the "source speaker" is a synthesized speech. For the training, target data can be extracted from the surrounding region of the gap or, in case of a famous speaker, in one embodiment it can be retrieved from a database, e.g. on the Internet. In another embodiment, training data for the target speaker can be recorded by e.g. asking the target

speaker to say some words, utterances or sentences. Then source data are synthesized with a text-to-speech synthesizer thanks to the transcript of the source speech, in one embodiment.

**[0010]** In one embodiment, text is extracted from the available speech signal by means of automatic speech recognition (ASR). Then, it is determined that one or more words or sounds are missing due to a gap in the speech signal, a context of the remainder of the speech signal is analyzed, and, according to the context and the remainder, one or more words, sounds or syllables are determined that are omitted by the gap. This can be done by estimating or guessing (e.g., in one embodiment by using a dictionary), or by obtaining from any source a complete transcript of the speech signal that covers at least the gap. It is easier to locate the gap if the complete transcript covers some more speech before and/or after the gap.

**[0011]** All following occurrences of the words "embodiment(s)" and "implementation(s)", if referring to feature combinations different from those defined by the independent claims, refer to examples which were originally filed but which do not represent embodiments/implementations of the presently claimed invention; these examples are still shown for illustrative purposes only.

**[0012]** In one embodiment, a computer readable medium has stored thereon executable instructions that when executed on a processor cause a processor to perform a method as disclosed above.

**[0013]** It is clear that in case of fully missing words it may in general simply be impossible to recover the missing speech, because it is not known what was said. At least some embodiments of the present principles provide a solution for this case, for example by generating the transcript based on the undistorted speech signal.

**[0014]** In one embodiment, a method for performing speech inpainting on a speech signal comprises automatically generating a transcript on an input speech signal, determining voice characteristics of the input speech signal, processing the input speech signal, whereby a processed speech signal is obtained, detecting a gap in the processed speech signal, automatically synthesizing from the transcript speech at least for the gap, voice converting the synthesized speech according to the determined voice characteristics, and inpainting the processed speech signal, wherein the voice converted synthesized speech is filled into the gap.

**[0015]** In one embodiment, an apparatus for performing speech inpainting on a speech signal comprises at least one hardware component, such as a hardware processor, and a non-transitory, tangible, computer-readable storage medium tangibly embodying at least one software component, and when executing on the at least one hardware processor, the software component causes the hardware processor to automatically perform the steps of claim 1.

**[0016]** Advantageous embodiments of the invention are disclosed in the dependent claims, the following description and the figures.

## Brief description of the drawings

**[0017]** Exemplary embodiments of the invention are described with reference to the accompanying drawings, which show in

- Fig.1 a general workflow of a speech inpainting system;
- Fig.2 embodiments of a voice conversion system;
- Fig.3 two embodiments for learning voice conversion,
- Fig.4 an overview of post-processing of the converted utterance,
- Fig.5 a flow-chart of a method for performing speech inpainting, and
- Fig.6 a block diagram of an apparatus for performing speech inpainting.

## Detailed description of the invention

**[0018]** Fig.1 shows a general workflow of a speech inpainting system. An input speech signal has a missing part 10, ie. a gap. A textual transcript of the missing part is available, for example it can be generated from the original speech signal. A speech utterance corresponding to the missing part 10 is synthesized 51 from the known text transcription through text-to-speech synthesis in a TTS synthesis block 11. However, such TTS synthesis systems may synthesize speech only phoneme by phoneme. Thus, if gaps occur in the middle of a phoneme, it is unlikely to recover only the utterance corresponding to the missing part. In one embodiment, it is more appropriate to synthesize as well the first and the last phoneme, corresponding to the beginning and the end of the missing part respectively, to reproduce the linguistic information because of the pronunciation context. It is also advantageous because it avoids speech discontinuities issues. After automatic speech synthesis 11, the generated speech is used for the gap filling. However, the synthesized speech has generally few similarities with the original speaker in terms of timbre and prosody. Therefore, its spectral features and fundamental frequency (F0) trajectory are adapted via voice conversion 12 to be similar to those of the target speech. Finally, the gap is filled by the voice converted synthesized speech signal, which results in an inpainted output signal 13. In the general pipeline, a conventional speech analysis-synthesis system (e.g.[6]) is used. This system enables performing flexible modifications on speech signals without loss of naturalness. In one embodiment, three parameters are extracted from the input signal: a STRAIGHT smooth spectrogram representing the evolution of the vocal tract without time and frequency interference, an F0 trajectory and a voice/unvoiced detector, and an aperiodic component. The first two parameters are manipulated by voice conversion to modify the speech. A STRAIGHT smooth spectrogram is known e.g.

from [6]. STRAIGHT is a speech tool for speech analysis and synthesis. It allows flexible manipulations on speech because it decomposes speech in the source-filter model in three parts: a smooth spectrum representing a spectral envelope, a fundamental frequency F0 measurement, and an aperiodic component. Basically, the fundamental frequency F0 measurement and the aperiodic component correspond to the source of the source-filter model, while the smooth spectrum representing a spectral envelope corresponds to the filter. The smooth STRAIGHT spectrum is a good representation of the envelope, because STRAIGHT reconstructs the envelope as if it was sampled by the source. Manipulating this spectrum allows us to make good modification of the timbre of the voice.

**[0019]** In one embodiment, the voice conversion system 12 comprises two steps. First a mapping function is learned on training data, and then it is used to convert new utterances. In order to get the mapping function, parameters to convert are extracted (e.g. with the STRAIGHT system) and aligned with dynamic time warping (DTW [8]). Then the learning phase is performed e.g. with a Gaussian mixture model (GMM [9]) or nonnegative matrix factorization (NMF [10])) to get the mapping function.

**[0020]** Fig.2 shows different embodiments of a voice conversion system, using a speech database. It is important to note that the original speech samples from the database do not necessarily need to cover the words or context of the current speech signal on which the inpainting is performed. The mapping function allowing to perform the prediction comprises two kind of parameters: general parameters that need to be calculated only once and parameters specific to the utterance that should be calculated for each utterance that is possible to convert. The general parameters may comprise e.g. Gaussian Mixture Model (GMM) parameters for GMM-based voice conversion and/or a phoneme dictionary for Non-negative Matrix Factorization (NMF)-based voice conversion. The specific parameters may comprise posterior probabilities for GMM-based voice conversion and/or temporal activation matrices for NMF-based voice conversion.

**[0021]** In one embodiment, where the speaker is a well-known person for whom many original speech samples can be retrieved from the Internet, the user is asked to enter, for a partly available speech signal 22, the speaker's identity in a query 21. The query results in voice characteristics of the speaker, or in original speech samples of the speaker from which the voice characteristics are extracted. The synthesized or original speech samples 23 obtained from a database or from the Internet 24 may be used to fill the gap. This approach may use standard voice conversion 25.

**[0022]** In another embodiment, where it is not possible to obtain sufficient original speech samples (e.g. because the speaker is not famous), voice characteristics of the speaker are retrieved upon a query 26 or automatically from the remaining part of the speech signal 27, which serve as a small set of training data. The synthesized speech for the gap 28 is voice converted 29 using the retrieved voice characteristics from around the gap.

**[0023]** Thus, two options may be considered to obtain training data, depending on whether the target speaker is a famous person or not. If the target speaker is e.g. well-known, it is generally possible to retrieve characteristic voice data from the Internet via the speaker's identity, or to try guessing the identity with automatic speaker recognition. Otherwise, only local data, i.e. data around the gap or some additional data, are available and the voice conversion system is adapted to the amount of data.

**[0024]** Fig.3 shows two embodiments 30,35 for learning voice conversion. As described above, a mapping function is learned on training data, and then it is used to convert new utterances. In order to get the mapping function 34, speech is generated from the training data by a text-to-speech block 31,38 (e.g. a speech synthesizer) and voice conversion parameters are extracted (e.g. with the STRAIGHT system) and aligned 32 to the synthesized speech with dynamic time warping (DTW). Then the learning phase is performed 33,39, e.g. with a Gaussian mixture model (GMM [9]) or Non-negative Matrix Factorization (NMF [10])), to get the mapping function 34. In one embodiment 30, only a small amount of training data is available, since only the speech surrounding the gap can be used as reliable speech to extract voice parameters. In another embodiment 33, a large amount of training data can be obtained from a database 36 such as the Internet, and automatic speech recognition 37 is used.

**[0025]** After speech parameters are converted thanks to the mapping function 34, a waveform signal is resynthesized, e.g. by a conventional STRAIGHT synthesizer with the new voice parameters.

**[0026]** In some embodiments, one or more additional steps may need to be performed, since once conversion is performed the resulting speech may still not perfectly fill the gap for the following reasons. First, edge mismatches such as spectral, fundamental frequency and phase discontinuities may need to be counteracted. Indeed, spectral trajectories of the formants are naturally smooth due to the slow variation of the vocal tract shape. Fundamental frequency and temporal phase are not as smooth as the spectral trajectories, but still need continuity to sound natural. Although the speech signal is converted, it is unlikely that the parameters of the spectral envelope trajectory, fundamental frequency and temporal phase are temporally continuous at the border of the gap. Thus, in one embodiment, the parameters of the spectral envelope trajectory, fundamental frequency and temporal phase are adapted to the ones nearby in the non-missing part of the speech, so that any discontinuity at the border is reduced. Besides, duration of the converted utterance may be longer or shorter than the true missing utterance. Therefore, in one embodiment, the speaking rate is converted to match the available part of the speech signal. If the speaking rate cannot be con-

verted, at least a temporal adjustment may be done on the global time scaling of the converted utterance.

**[0027]** A method dealing with these issues is briefly outlined in Fig.4, which shows an overview of post-processing of the converted utterance. First, the converted set of frames may not properly fill the gap. This can be seen as spectral discontinuities 4a. According to an embodiment of the present principles, the gaps may be properly filled by finding 41 in the spectral domain the best frames at the end of the converted spectrogram and merging them with the reliable spectrogram of the available portion of speech signal. This can be done by the known dynamic time warping (DTW) algorithm. Aligning converted and reliable spectra is a way to find which data are used to fill the gap. Then, in a similar adjustment to handle phase discontinuities 4b, the best samples to merge are found 42 on the signal waveform. Such issue appears when for instance speech is voice converted and the waveform signal has the particularity to be periodic. This property is used in cross-correlation between the edges of the reliable signal and the beginning of the converted signal. Peaks in the cross-correlation point out best indices to merge both signals. Then, a fundamental frequency F0 trajectory is modified 43 so that F0 and F0 derivative ($d$F0/$d$t) discontinuities 4c are minimized especially on the edges of the converted signal 4d. The F0 trajectory can be computed in the same way as for spectral parameters. The edges of the resulting signal are "allocated" to gap edges. However, the body of the signal may not be suited to the gap: it may be too large or too small. Therefore, in one embodiment the converted signal with F0 modification is time-scaled 44 (without pitch modification, in one embodiment) according to the indices found in the phase adjustment step. Finally, the length-adjusted (ie. "stretched" or "compressed") signal is overlap-added 45 on the edges to minimize fuzzy artefacts that could still remain.

**[0028]** One advantage of the disclosed audio inpainting technique is that even long gaps can be inpainted. It is also robust when only a small amount of data is available in voice conversion.

**[0029]** Fig.5 shows a flow-chart of a method for performing speech inpainting on a speech signal, according to one embodiment. The method 50 comprises determining 51 voice characteristics of the speech signal, detecting 52 a gap in the speech signal, automatically synthesizing 53, from a transcript, speech at least for the gap, voice converting 54 the synthesized speech according to the determined voice characteristics, and inpainting 55 the speech signal, wherein the voice converted synthesized speech is filled into the gap.

**[0030]** In one embodiment, the method further comprises a step of automatically generating 56 said transcript on an input speech signal.

**[0031]** In one embodiment, the method further comprises a step of processing 57 the voice signal, wherein the gap is generated during the processing, and wherein the transcript is generated before the processing.

**[0032]** In one embodiment, the step of automatically synthesizing 53 speech at least for the gap comprises retrieving from a database recorded speech data from a natural speaker. This may support, enhance, replace or control the synthesis.

**[0033]** In one embodiment, the method further comprises steps of detecting 581 that the transcript does not cover the gap, determining 582 one or more words or sounds omitted by the gap, and adding 583 the estimated word or sound to the transcript before synthesizing speech from the transcript.

**[0034]** In one embodiment, the determining 582 is done by estimating or guessing the one or more words or sounds (e.g. from a dictionary).

**[0035]** In one embodiment, the determining 582 is done by retrieving a complete transcript of the speech through other channels (e.g. the Internet).

**[0036]** In one embodiment, the determined voice characteristics comprise parameters for a spectral envelope and a fundamental frequency F0 (or, in other words, it is timbre and prosody of the speech).

**[0037]** In one embodiment, the method further comprises adapting parameters for a spectral envelope trajectory, a fundamental frequency and temporal phase at one or both boundaries of the gap to match the corresponding parameters of the available adjacent speech signal before and/or after the gap. This is in order for the parameters to be temporally continuous before and/or after the gap.

**[0038]** In one embodiment, the method further comprises a step of time-scaling the voice-converted speech signal before it is filled into the gap.

**[0039]** Fig.6 shows a block diagram of an apparatus 60 for performing speech inpainting on a speech signal, according to one embodiment. The apparatus comprises a speech analyser 61 for detecting a gap G in the speech signal SI, a speech synthesizer 62 for automatically synthesizing from a transcript T speech SS at least for the gap, a voice converter 63 for converting the synthesized speech SS according to the determined voice characteristics VC, and a mixer 64 for inpainting the speech signal, wherein the voice converted synthesized speech VCS is filled into the gap G of the speech signal to obtain an inpainted speech output signal SO.

**[0040]** In one embodiment, the apparatus further comprises a voice analyzer 65 for determining voice characteristics of the speech signal.

**[0041]** In one embodiment, the apparatus further comprises a speech-to-text converter 66 for automatically generating a transcript of the speech signal.

**[0042]** In one embodiment, the apparatus further comprises a database having stored speech data of example phonemes or words of natural speech, and the speech synthesizer 62 retrieves speech data from the database for automatically synthesizing the speech at least for the gap.

**[0043]** In one embodiment, the apparatus further comprises an interface 67 for receiving a complete transcript

of the speech signal, the transcript covering at least text that is omitted by the gap.

**[0044]** In one embodiment, the apparatus further comprises a time-scaler for time-scaling the voice-converted speech signal before it is filled into the gap.

**[0045]** In one embodiment, an apparatus for performing speech inpainting on a speech signal comprises a processor and a memory storing instructions that, when executed by the processor, cause the apparatus to perform the method steps of any of the methods disclosed above.

**[0046]** It is noted that the use of the verb "comprise" and its conjugations does not exclude the presence of elements or steps other than those stated in a claim. Furthermore, the use of the article "a" or "an" preceding an element does not exclude the presence of a plurality of such elements. Furthermore, the invention resides in each and every novel feature or combination of features.

**[0047]** It should be noted that although the STRAIGHT system is mentioned, other types of speech analysis and synthesis systems may be used other than STRAIGHT, as would be apparent to those of ordinary skill in the art, all of which are contemplated within the scope of the invention.

**[0048]** While there has been shown, described, and pointed out fundamental novel features of the present invention as applied to preferred embodiments thereof, it will be understood that various omissions and substitutions and changes in the apparatus and method described, in the form and details of the devices disclosed, and in their operation, may be made by those skilled in the art without departing from the scope of the present invention. It is expressly intended that all combinations of those elements that perform substantially the same function in substantially the same way to achieve the same results are within the scope of the invention. Substitutions of elements from one described embodiment to another are also fully intended and contemplated. Each feature disclosed in the description and (where appropriate) the claims and drawings may be provided independently or in any appropriate combination. Features may, where appropriate be implemented in hardware, software, or a combination of the two. Reference numerals appearing in the claims are by way of illustration only and shall have no limiting effect on the scope of the claims.

**[0049]** The scope of the present invention is defined in the appended claims.

Cited References

**[0050]**

[1] Amir Adler, Valentin Emiya, Maria Jafari, Michael Elad, Remi Gribonval, Mark D. Plumbley, "Audio inpainting," IEEE Transactions on Audio, Speech and Language Processing, IEEE, 2012, 20 (3), pp. 922 - 932 , XP011397627

[2] P. Smaragdis et al. "Missing data imputation for spectral audio signal," Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2009

[3] J. Le Roux et al., "Computational auditory induction as a missing data model-fitting problem with Bregman divergence," Speech Communication, vol. 53, no. 5, pp. 658-676, 2011

[4] I. Drori et al. "Spectral sound gap filling," Proc. ICPR 2004, pp. 871-874

[5] Jani Nurminen, Hanna Silen, Victor Popa, Elina Helander and Moncef Gabbouj (2012). "Voice Conversion, Speech Enhancement, Modeling and Recognition- Algorithms and Applications", Dr. S Ramakrishnan (Ed.), ISBN: 978-953-51-0291-5, InTech, DOI: 10.5772/37334.

[6] Hideki Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In 1997 IEEE Inter- national Conference on Acoustics, Speech, and Signal Processing, ICASSP '97, Munich, Germany, April 21-24, 1997, pages 1303-1306, 1997.

[7] Y. Bahat, Y. Y. Schechner, and M. Elad, "Self-content-based audio inpainting," Signal Processing, vol. 111, pp. 61-72, 2015.

[8] D. Ellis (2003). Dynamic Time Warp (DTW) in Matlab, Web resource, available at http://www.ee.columbia.edu/-dpwe/resources/matlab/dtw/. Visited 4/29/2015.

[9] Toda, T.; Black, A.W.; Tokuda, K., "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," Audio, Speech, and Language Processing, IEEE Transactions on, vol.15, no.8, pp.2222,2235, Nov. 2007

[10] Aihara, R.; Nakashika, T.; Takiguchi, T.; Ariki, Y., "Voice conversion based on Non-negative matrix factorization using phoneme-categorized dictionary," Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on , vol., no., pp.7894,7898, 4-9 May 2014

## Claims

**1.** A method (50) comprising:

- obtaining (51) voice characteristics of a speech signal;
- detecting (52) a gap in the speech signal;
- automatically synthesizing (53), from a transcript, speech at least for the gap in the speech signal;
- voice converting (54) the synthesized speech according to the obtained voice characteristics of the speech signal; and
- inpainting (55) the speech signal, wherein the voice converted synthesized speech is filled into the gap.

2. The method of claim 1, comprising automatically generating (56) said transcript from the speech signal.

3. The method according to claim 1 or 2, comprising processing (57) the speech signal, wherein the gap occurs during the processing, and wherein the transcript is generated before the processing.

4. The method according to any of claims 1-3, wherein automatically synthesizing (53), from a transcript, speech at least for the gap comprises retrieving from a database recorded speech data from a natural speaker.

5. The method according to any of claims 1-4, comprising

- detecting (581) that the transcript does not cover the gap;
- determining (582) one or more words or sounds omitted within the gap; and
- adding (583) the determined word or sound to the transcript before synthesizing speech from the transcript.

6. The method according to claim 5, wherein the determining (582) is done by estimating or guessing the one or more words or sounds.

7. The method according to claim 5, wherein the determining (582) is done by retrieving a complete transcript of the speech through other channels.

8. The method according to any of claims 1-7, wherein the voice characteristics comprise parameters for a spectral envelope and a fundamental frequency.

9. The method according to one of the claims 1-8, comprises adapting parameters for a spectral envelope trajectory, a fundamental frequency and a temporal phase at one or both boundaries of the gap to match the corresponding parameters of the available adjacent speech signal before and/or after the gap.

10. The method according to one of the claims 1-9, comprising time-scaling the voice-converted speech signal before it is filled into the gap.

11. An apparatus (60) comprising:

- a speech analyser (61) for detecting gap in a speech signal;
- a speech synthesizer (62) for automatically synthesizing, from a transcript, speech at least for a gap in the speech signal;
- means for obtaining voice characteristics of the speech signal;
- a voice converter (63) for converting the synthesized speech according to the obtained voice characteristics of the speech signal; and
- a mixer (64) for inpainting the speech signal, wherein the voice converted synthesized speech is filled into the gap of the speech signal.

12. The apparatus of claim 11, wherein said means for obtaining comprises a voice analyzer (65) for obtaining the voice characteristics of the speech signal.

13. The apparatus of claim 11 or 12, comprising a speech-to-text converter (66) for automatically generating a transcript of the speech signal.

14. The apparatus of one of the claims 11-13, comprising a database having stored speech data of example phonemes or words of natural speech, wherein the speech synthesizer (62) retrieves speech data from the database for automatically synthesizing the speech at least for the gap.

15. The apparatus of one of the claims 11-14, comprising an interface (67) for receiving a complete transcript of the speech signal, the transcript covering at least text that is omitted by the gap.

**Patentansprüche**

1. Verfahren (50), das umfasst:

- Erhalten (51) einer Stimmcharakteristik eines Sprachsignals;
- Detektieren (52) einer Lücke in dem Sprachsignal;
- automatisches Synthetisieren (53) von Sprache mindestens für die Lücke in dem Sprachsignal aus einer Niederschrift;
- Umsetzen (54) der Stimme der synthetisierten Sprache gemäß der erhaltenen Stimmcharakteristik des Sprachsignals; und
- Inpainting (55) des Sprachsignals, wobei die synthetisierte Sprache mit umgesetzter Stimme in die Lücke gefüllt wird.

2. Verfahren nach Anspruch 1, das das automatische Erzeugen (56) der Niederschrift aus dem Sprachsignal umfasst.

3. Verfahren nach Anspruch 1 oder 2, das das Verarbeiten (57) des Sprachsignals umfasst, wobei die Lücke während der Verarbeitung auftritt und wobei die Niederschrift vor der Verarbeitung erzeugt wird.

4. Verfahren nach einem der Ansprüche 1-3, wobei das automatische Synthetisieren (53) von Sprache mindestens für die Lücke aus einer Niederschrift das Auslesen aufgezeichneter Sprachdaten von einem natürlichen Sprecher aus einer Datenbank umfasst.

5. Verfahren nach einem der Ansprüche 1-4, das umfasst:

- Detektieren (581), dass die Niederschrift die Lücke nicht abdeckt;
- Bestimmen (582) eines oder mehrerer Wörter oder Schalle, die in der Lücke weggelassen sind; und
- Hinzufügen (583) des bestimmten Worts oder Schalls zu der Niederschrift, bevor die Sprache aus der Niederschrift synthetisiert wird.

6. Verfahren nach Anspruch 5, wobei das Bestimmen (582) durch Schätzen oder Vermuten des einen oder der mehreren Wörter oder Schalle erfolgt.

7. Verfahren nach Anspruch 5, wobei das Bestimmen (582) durch Auslesen einer vollständigen Niederschrift der Sprache über andere Kanäle erfolgt.

8. Verfahren nach einem der Ansprüche 1-7, wobei die Stimmcharakteristiken Parameter für eine spektrale Einhüllende und für eine Grundfrequenz umfassen.

9. Verfahren nach einem der Ansprüche 1-8, das das Anpassen von Parametern für eine Trajektorie der spektralen Einhüllenden, für eine Grundfrequenz und für eine zeitliche Phase bei einer oder beiden Grenzen der Lücke umfasst, um die entsprechenden Parameter des verfügbaren angrenzenden Sprachsignals vor und/oder nach der Lücke anzupassen.

10. Verfahren nach einem der Ansprüche 1-9, das das zeitliche Skalieren des Sprachsignals mit umgesetzter Stimme, bevor es in die Lücke gefüllt wird, umfasst.

11. Vorrichtung (60), die umfasst:

- einen Sprachanalysator (61) zum Detektieren einer Lücke in einem Sprachsignal;
- einen Sprachsynthetisator (62) zum automatischen Synthetisieren von Sprache mindestens für eine Lücke in dem Sprachsignal aus einer Niederschrift;
- ein Mittel zum Erhalten von Stimmcharakteristiken des Sprachsignals;
- einen Stimmumsetzer (63) zum Umsetzen der synthetisierten Sprache in Übereinstimmung mit den erhaltenen Stimmcharakteristiken des Sprachsignals; und
- einen Mischer (64) für das Inpainting des Sprachsignals, wobei die synthetisierte Sprache mit umgesetzter Stimme in die Lücke des Sprachsignals gefüllt wird.

12. Vorrichtung nach Anspruch 11, wobei das Mittel zum Erhalten einen Stimmanalysator (65) zum Erhalten der Stimmcharakteristiken des Sprachsignals umfasst.

13. Vorrichtung nach Anspruch 11 oder 12, die einen Sprache-zu-Text-Umsetzer (66) zum automatischen Erzeugen einer Niederschrift des Sprachsignals umfasst.

14. Vorrichtung nach einem der Ansprüche 11-13, die eine Datenbank umfasst, in der Sprachdaten von Beispielphonemen oder -wörtern natürlicher Sprache gespeichert sind, wobei der Sprachsynthetisator (62) Sprachdaten aus der Datenbank ausliest, um die Sprache mindestens für die Lücke automatisch zu synthetisieren.

15. Vorrichtung nach einem der Ansprüche 11-14, die eine Schnittstelle (67) zum Empfangen einer vollständigen Niederschrift des Sprachsignals umfasst, wobei die Niederschrift mindestens Text abdeckt, der durch die Lücke weggelassen ist.

**Revendications**

1. Procédé (50) comprenant :

- une obtention (51) des caractéristiques vocales d'un signal vocal ;
- une détection (52) d'une partie manquante dans le signal vocal ;
- une synthèse automatique (53), à partir d'une transcription, de parole au moins pour la partie manquante dans le signal vocal ;
- une conversion en voix (54) de la parole synthétisée selon les caractéristiques vocales obtenues du signal vocal ; et
- une retouche (55) du signal vocal, où la parole synthétisée convertie en voix est insérée dans la partie manquante.

2. Procédé selon la revendication 1, comprenant une génération automatique (56) de ladite transcription à partir du signal vocal.

3. Procédé selon la revendication 1 ou 2, comprenant un traitement (57) du signal vocal, dans lequel la partie manquante survient pendant le traitement et dans lequel la transcription est générée avant le traitement.

4. Procédé selon l'une quelconque des revendications

1 à 3, dans lequel la synthèse automatique (53), à partir d'une transcription, de la parole au moins pour la partie manquante comprend une extraction à partir d'une base de données de données vocales enregistrées par une voix humaine.

**5.** Procédé selon l'une quelconque des revendications 1 à 4, comprenant les étapes consistant à :

- détecter (581) que la transcription ne couvre pas la partie manquante ;
- déterminer (582) un ou plusieurs mots ou sons omis dans la partie manquante ; et
- ajouter (583) le mot ou le son déterminé à la transcription avant la synthèse de la parole à partir de la transcription.

**6.** Procédé selon la revendication 5, dans lequel la détermination (582) est effectuée en estimant ou en devinant le ou les mots ou sons.

**7.** Procédé selon la revendication 5, dans lequel la détermination (582) s'effectue en extrayant une transcription complète de la parole via d'autres canaux.

**8.** Procédé selon l'une quelconque des revendications 1 à 7, dans lequel les caractéristiques vocales comprennent des paramètres pour une enveloppe spectrale et une fréquence fondamentale.

**9.** Procédé selon l'une des revendications 1 à 8, comprenant une adaptation de paramètres pour une trajectoire d'enveloppe spectrale, une fréquence fondamentale et une phase temporelle au niveau d'une ou des deux limites de la partie manquante afin d'établir une correspondance avec les paramètres correspondants du signal vocal adjacent disponible avant et/ou après la partie manquante.

**10.** Procédé selon l'une des revendications 1 à 9, comprenant une mise à l'échelle temporelle du signal vocal converti en voix avant son insertion dans la partie manquante.

**11.** Appareil (60) comprenant :

- un analyseur vocal (61) pour détecter une partie manquante dans un signal vocal ;
- un synthétiseur vocal (62) pour synthétiser automatiquement, à partir d'une transcription, une parole au moins pour une partie manquante dans le signal vocal ;
- un moyen pour obtenir les caractéristiques vocales du signal vocal :
- un convertisseur vocal (63) pour convertir la parole synthétisée selon les caractéristiques vocales obtenues du signal vocal ; et
- un mélangeur (64) pour retoucher le signal vocal, où la parole synthétisée convertie en voix est insérée dans la partie manquante du signal vocal.

**12.** Appareil selon la revendication 11, dans lequel ledit moyen d'obtention comprend un analyseur vocal (65) pour obtenir les caractéristiques vocales du signal vocal.

**13.** Appareil selon la revendication 11 ou 12, comprenant un convertisseur voix-texte (66) pour générer automatiquement une transcription du signal vocal.

**14.** Appareil selon l'une des revendications 11 à 13, comprenant une base de données contenant des données vocales d'exemples de phonèmes ou de mots de voix humaine, dans lequel le synthétiseur vocal (62) extrait des données vocales de la base de données pour synthétiser automatiquement la parole au moins pour la partie manquante.

**15.** Appareil selon l'une des revendications 11 à 14, comprenant une interface (67) pour recevoir une transcription complète du signal vocal, la transcription couvrant au moins le texte omis par la partie manquante.

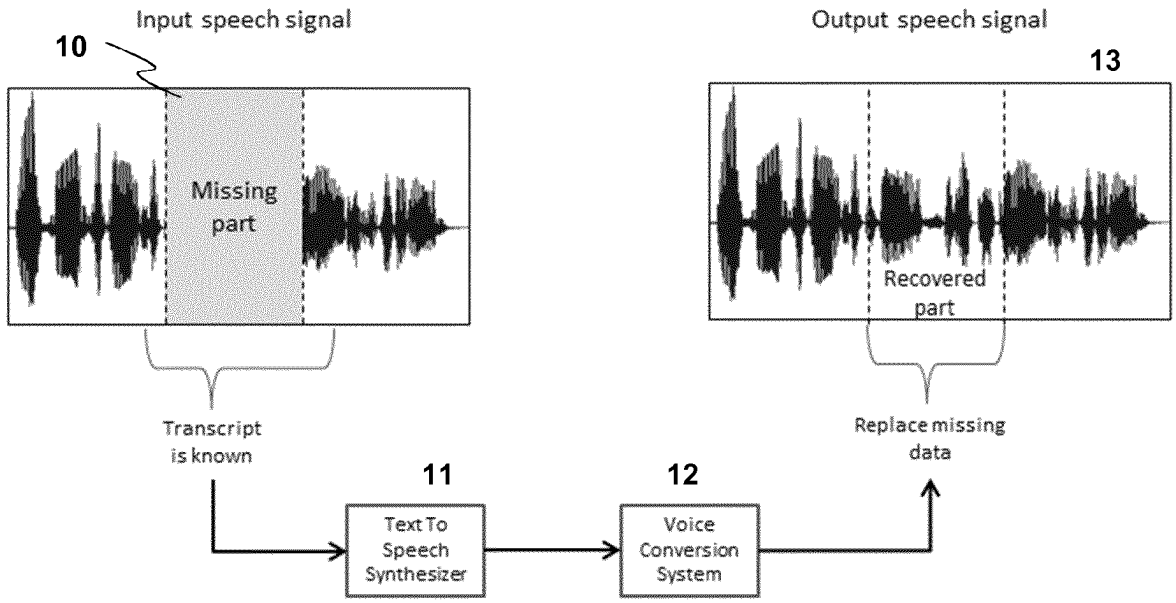
Input speech signal
10
Missing part
Output speech signal
13
Recovered part
Transcript is known
Replace missing data
11
12
Text To Speech Synthesizer
Voice Conversion System

Fig.1

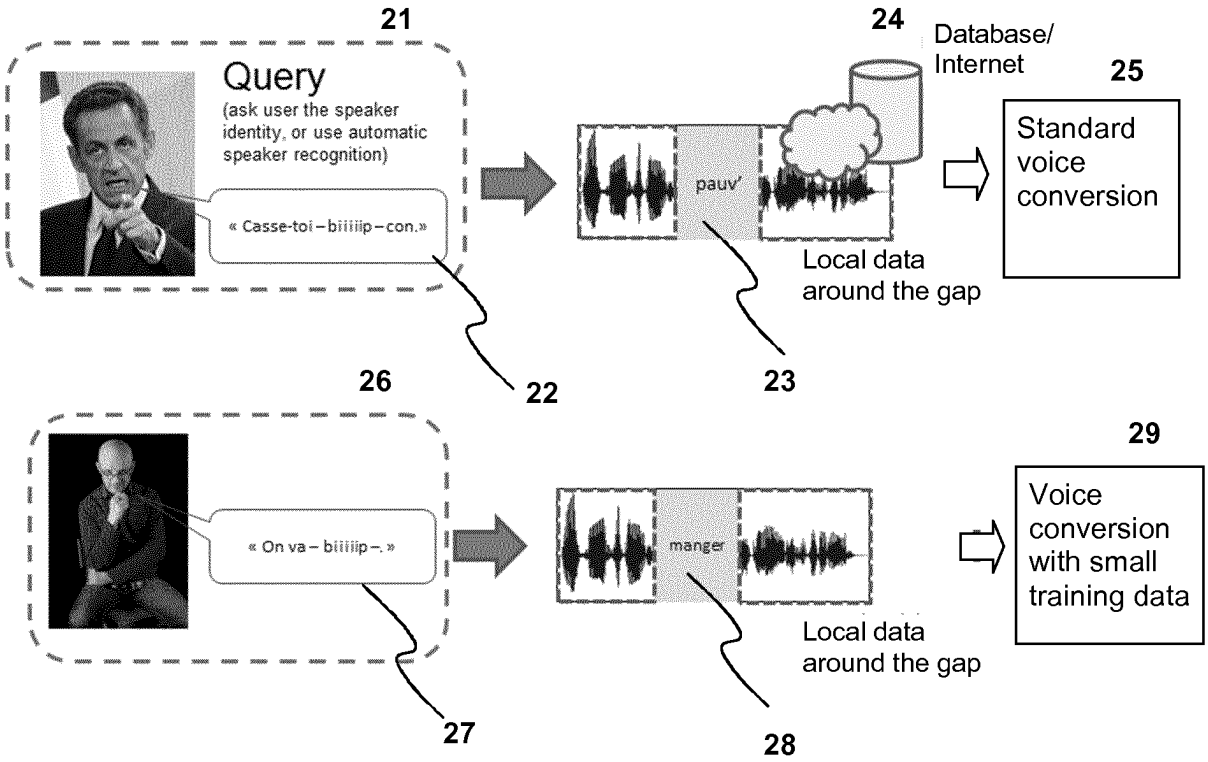
21
Query
(ask user the speaker identity, or use automatic speaker recognition)
« Casse-toi – biiiiip – con.»
22
24
Database/ Internet
25
Standard voice conversion
pauv'
Local data around the gap
23
26
« On va – biiiiip –. »
27
29
Voice conversion with small training data
manger
Local data around the gap
28

Fig.2

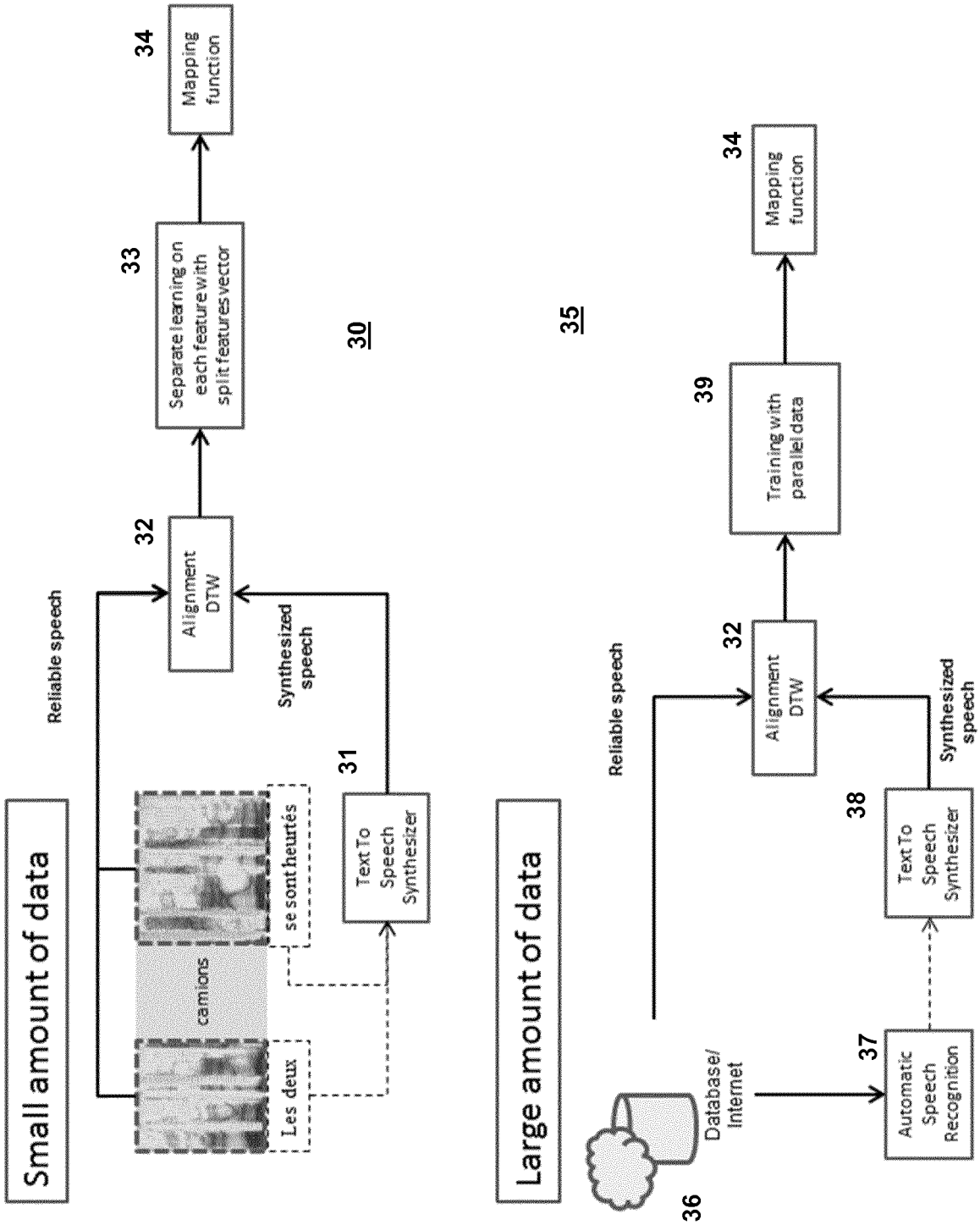
Small amount of data
Reliable speech
Les deux
camions
se sont heurtés
Text To Speech Synthesizer
31
Synthesized speech
Alignment DTW
32
Separate learning on each feature with split features vector
33
Mapping function
34
30
Large amount of data
Reliable speech
Database/ Internet
36
Automatic Speech Recognition
37
Text To Speech Synthesizer
38
Synthesized speech
Alignment DTW
32
Training with parallel data
39
Mapping function
34
35

Fig.3

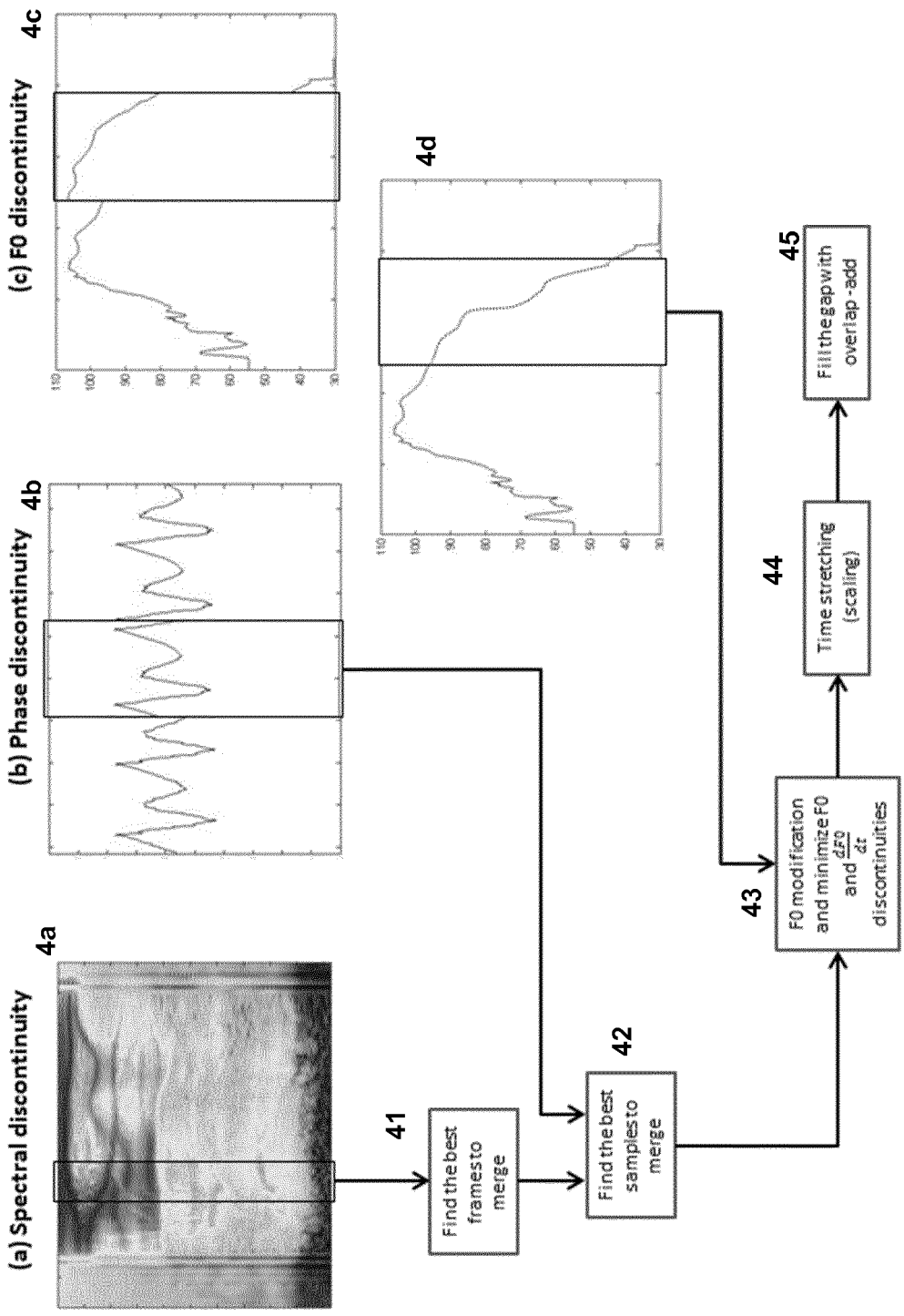
(a) Spectral discontinuity
4a
(b) Phase discontinuity
4b
(c) F0 discontinuity
4c
4d
41
Find the best frames to merge
42
Find the best samples to merge
43
F0 modification and minimize F0 and $\frac{dF0}{dt}$ discontinuities
44
Time stretching (scaling)
45
Fill the gap with overlap-add

Fig.4

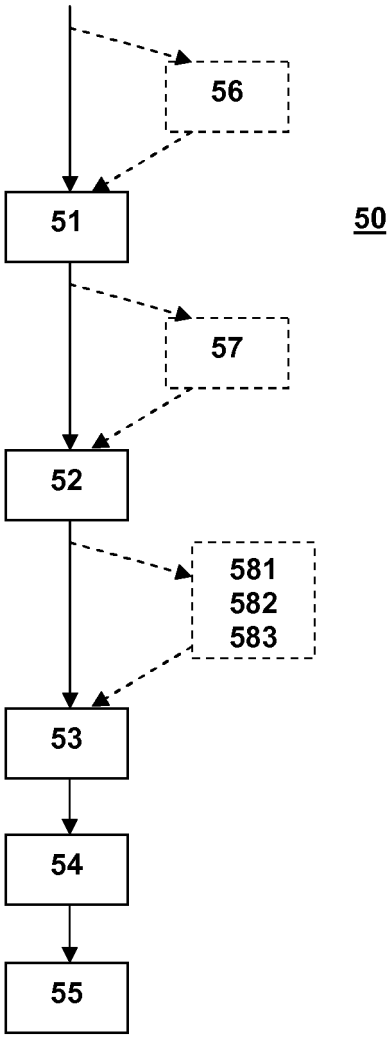
56
51
50
57
52
581
582
583
53
54
55

Fig.5

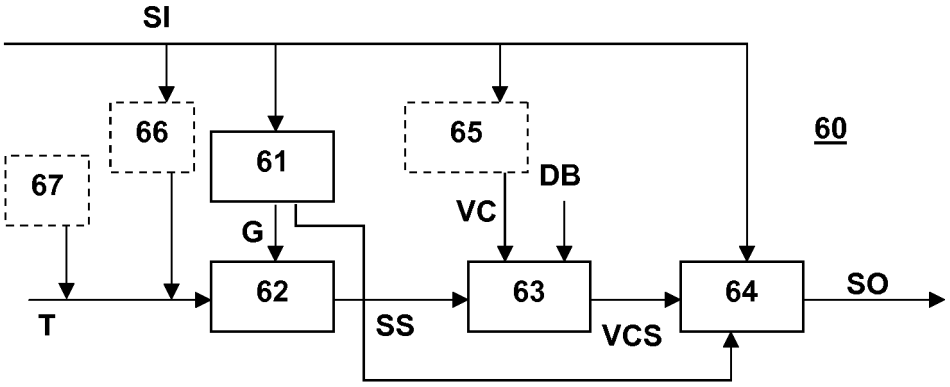
SI
66
61
65
60
67
DB
VC
G
62
63
64
SO
T
SS
VCS

Fig.6

## REFERENCES CITED IN THE DESCRIPTION

*This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.*

### Non-patent literature cited in the description


- Audio inpainting. **AMIR ADLER ; VALENTIN EMIYA ; MARIA JAFARI ; MICHAEL ELAD ; REMI GRIBONVAL ; MARK D. PLUMBLEY.** IEEE Transactions on Audio, Speech and Language Processing. IEEE, 2012, vol. 20, 922-932 **[0050]**
- **P. SMARAGDIS et al.** Missing data imputation for spectral audio signal. *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP),* 2009 **[0050]**
- **J. LE ROUX et al.** Computational auditory induction as a missing data model-fitting problem with Bregman divergence. *Speech Communication,* 2011, vol. 53 (5), 658-676 **[0050]**
- **I. DRORI et al.** Spectral sound gap filling. *Proc. ICPR,* 2004, 871-874 **[0050]**
- **JANI NURMINEN ; HANNA SILEN ; VICTOR POPA ; ELINA HELANDER ; MONCEF GABBOUJ.** Voice Conversion, Speech Enhancement, Modeling and Recognition- Algorithms and Applications. 2012 **[0050]**
- **HIDEKI KAWAHARA.** Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. *In 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '97,* 1997, 1303-1306 **[0050]**
- **Y. BAHAT ; Y. Y. SCHECHNER ; M. ELAD.** Self-content-based audio inpainting. *Signal Processing,* 2015, vol. 111, 61-72 **[0050]**
- **D. ELLIS.** Dynamic Time Warp (DTW). *Matlab, Web resource,* 2003, http://www.ee.columbia.edu/-dpwe/resources/matlab/dtw **[0050]**
- **TODA, T. ; BLACK, A.W. ; TOKUDA, K.** Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on,* November 2007, vol. 15 (8), 2222, , 2235 **[0050]**
- **AIHARA, R. ; NAKASHIKA, T. ; TAKIGUCHI, T. ; ARIKI, Y.** Voice conversion based on Non-negative matrix factorization using phoneme-categorized dictionary. *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference,* 04 May 2014, 7894, , 7898 **[0050]**