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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 September 2020 (03.09.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/172831 A1

(51) International Patent Classification:
G10L 21/0272 (2013.01)

(21) International Application Number:
PCT/CN2019/076395

(22) International Filing Date:
28 February 2019 (28.02.2019)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **BEIJING DIDI INFINITY TECHNOLOGY AND DEVELOPMENT CO., LTD.** [CN/CN]; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN).

(72) Inventors: **ZHANG, Yi**; 450 National Ave., Mountain View, California 94043 (US). **SONG, Hui**; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN). **SHA, Yongtao**; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN). **DENG, Chengyun**; Building 34, No. 8 Dongbeiwang West Road, Haidian District, Beijing 100193 (CN).

(74) Agent: **METIS IP (CHENGDU) LLC**; Room 808, 8th Floor, Building B7, Block D, Tianfu Jingrong Center, No. 99 West Section of Hupan Road, Tianfu New District, Chengdu, Sichuan 610213 (CN).

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, 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).

Published:
— *with international search report (Art. 21(3))*

(54) Title: CONCURRENT MULTI-PATH PROCESSING OF AUDIO SIGNALS FOR AUTOMATIC SPEECH RECOGNITION SYSTEMS

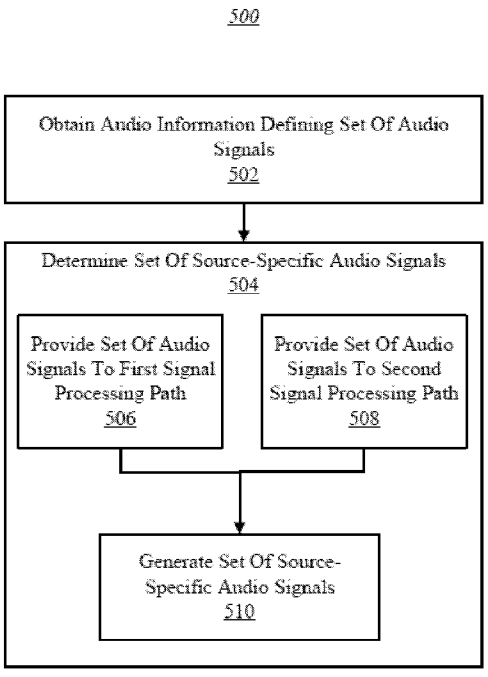

FIG. 5

(57) Abstract: A system and method for concurrent multi-path processing of audio signals for automatic speech recognition is presented. Audio information defining a set of audio signals may be obtained (502). The audio signals may convey mixed audio content produced by multiple audio sources. A set of source-specific audio signals may be determined by demixing the mixed audio content produced by the multiple audio sources. Determining the set of source-specific audio signals may comprises providing the set of audio signals to both a first signal processing path and a second signal processing path (504). The first signal processing path may determine a value of a demixing parameter for demixing the mixed audio content (506). The second signal processing path may apply the value of the demixing parameter to the individual audio signals of the set of audio signals (508) to generate the individual source-specific audio signals (510).

WO 2020/172831 A1

CONCURRENT MULTI-PATH PROCESSING OF AUDIO SIGNALS FOR AUTOMATIC SPEECH RECOGNITION SYSTEMS

TECHNICAL FIELD

[0001] The disclosure relates generally to concurrent multi-path processing of audio signals for automatic speech recognition systems.

BACKGROUND

[0002] Speech assistant systems utilizing automatic speech recognition (ASR) are widely used in cars and other vehicles. Given voice commands, the systems may recognize the instruction and/or operate a function automatically, *e.g.*, making phone calls, controlling AC, etc. Compared to manual operation, speech assistant systems keep drivers focusing on road condition, which greatly increases the driving safety.

SUMMARY

[0003] Automatic speech recognition (ASR) systems may yield good results in quiet environments, *e.g.*, word error rate (WER) may drop down to 5% or less. However, their performance drops dramatically when the desired speech is corrupted by interferences, including competing talkers, background noise, and so on. There still remains a big challenge making ASR robust in real life applications where the environment may be noisy. ASR works well in lab/quiet conditions. But in real life driving conditions, ambient noise, air conditioning, traffic noise, and error rate increase significantly – can be below 60% - which may not be useful.

[0004] A pre-processing stage, to suppress interference and distort desired speech at minimum, is often considered an efficient technique to improve ASR robustness. Generally, better performance can be achieved when more

microphones are used. But in practice small number of microphones (*e.g.*, 2 or 3 microphones) may be preferred in automobile applications. One or more aspects of the present disclosure propose systems and methods implementing a real-time blind source separation technique to provide improved automatic speech recognition based on concurrent multi-path processing of audio signals. The proposed systems and methods may achieve improved recognition accuracy even with the existence of one or a combination of interference, real-time algorithm delay, or a relatively small number of microphones (*e.g.*, 2 or 3).

[0005] One aspect of the present disclosure is directed to a method for concurrent multi-path processing of audio signals for automatic speech recognition. The method may comprise: obtaining audio information defining a set of audio signals, individual audio signals in the set of audio signals conveying mixed audio content produced by multiple audio sources, the mixed audio content including speech and noise; and determining a set of source-specific audio signals by demixing the mixed audio content produced by the multiple audio sources, individual source-specific audio signals representing individual audio content produced by specific individual audio sources of the multiple audio sources, wherein determining the set of source-specific audio signals comprises: providing the set of audio signals to a first signal processing path to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources; providing the set of audio signals to a second signal processing path to apply the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path; and generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals.

[0006] Another aspect of the present disclosure is directed to a system for concurrent multi-path processing of audio signals for automatic speech

recognition. The system may comprise one or more processors, a memory storing instructions, and a speech recognition engine. The instructions, when executed by the one or more processors, may cause the system to perform: obtaining audio information defining a set of audio signals, individual audio signals in the set of audio signals conveying mixed audio content produced by multiple audio sources, the mixed audio content including speech and noise; and determining a set of source-specific audio signals by demixing the mixed audio content produced by the multiple audio sources, individual source-specific audio signals representing individual audio content produced by specific individual audio sources of the multiple audio sources, wherein determining the set of source-specific audio signals comprises: providing the set of audio signals to a first signal processing path to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources; providing the set of audio signals to a second signal processing path to apply the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path; and generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals, such that a first source-specific audio signal represents the speech and a second source-specific audio signal represents the noise.

[0007] In some embodiments, providing the set of audio signals to the first signal processing path further includes the operations of: transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals; collecting the consecutive frames of the time-frequency domain representation of the individual audio signals into individual sets of combined consecutive frames of the time-frequency domain representations; and for the individual sets of combined consecutive frames of the time-frequency domain representations:

approximating individual reduced dimensionality representations of the individual sets of combined consecutive frames of the time-frequency domain representation of the individual audio signals; decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations; determining a current value of the demixing parameter based on the decomposed individual reduced dimensionality representations; comparing the current value of the demixing parameter to a previous value of the demixing parameter; and based on the comparison, setting the value of the demixing parameter as either the current value or a modified version of the current value.

[0008] In some embodiments, the operation of transforming the individual audio signals into consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

[0009] In some embodiments, the operation of approximating individual reduced dimensionality representations comprises a principal component analysis (PCA) whitening operation.

[0010] In some embodiments, providing the set of audio signals to the second signal processing path further includes the operations of: transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals; and consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals: approximating an individual reduced dimensionality representation of the time-frequency domain representation of an individual audio signal included in an individual frame; and applying the value of the demixing parameter to the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to

obtain an individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame.

[0011] In some embodiments, the operation of transforming the individual audio signals to consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

[0012] In some embodiments, the operation of approximating the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame comprises a principal component analysis (PCA) whitening operation.

[0013] In some embodiments, generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals comprises operations of: consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals: restoring the dimensionality of the individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed time-frequency domain representation of the individual audio signal included in the individual frame; and transforming the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to a time domain representation; and wherein the time domain representation of the consecutive frames of the individual demixed time-frequency domain representation of the individual audio signals comprises the individual source-specific audio signals.

[0014] In some embodiments, the operation of transforming the individual demixed time-frequency domain representation of the individual audio signal

included in the individual frame to the time domain representation comprises a synthesis filter bank (SFB) operation.

[0015] These and other features of the systems, methods, and non-transitory computer readable media disclosed herein, as well as the methods of operation and functions of the related elements of structure and the combination of parts and economies of manufacture, will become more apparent upon consideration of the following description and the appended claims with reference to the accompanying drawings, all of which form a part of this specification, wherein like reference numerals designate corresponding parts in the various figures. It is to be expressly understood, however, that the drawings are for purposes of illustration and description only and are not intended as a definition of the limits of the invention. It is to be understood that the foregoing general description and the following detailed description are exemplary and explanatory only, and are not restrictive of the invention, as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Preferred and non-limiting embodiments of the invention may be more readily understood by referring to the accompanying drawings in which:

[0017] FIG. 1 illustrates an example environment for concurrent multi-path processing of audio signals for automatic speech recognition systems, in accordance with various embodiments of the disclosure.

[0018] FIG. 2 an example environment for concurrent multi-path processing of audio signals for automatic speech recognition systems, in accordance with various embodiments of the disclosure.

[0019] FIG. 3 an example environment for concurrent multi-path processing of audio signals for automatic speech recognition systems, in accordance with various embodiments of the disclosure.

[0020] FIG. 4 an example environment for concurrent multi-path processing of audio signals for automatic speech recognition systems, in accordance with various embodiments of the disclosure.

[0021] FIG. 5 illustrates an example flow chart of concurrent multi-path processing of audio signals for automatic speech recognition, in accordance with various embodiments of the disclosure.

[0022] FIG. 6 illustrates an example flow chart of concurrent multi-path processing of audio signals for automatic speech recognition, in accordance with various embodiments of the disclosure.

[0023] FIG. 7 illustrates a block diagram of an example computer system in which any of the embodiments described herein may be implemented.

DETAILDE DESCRIPTION

[0024] Specific, non-limiting embodiments of the present invention will now be described with reference to the drawings. It should be understood that particular features and aspects of any embodiment disclosed herein may be used and/or combined with particular features and aspects of any other embodiment disclosed herein. It should also be understood that such embodiments are by way of example and are merely illustrative of a small number of embodiments within the scope of the present invention. Various changes and modifications obvious to one skilled in the art to which the present invention pertains are deemed to be within the spirit, scope and contemplation of the present invention as further defined in the appended claims.

[0025] The approaches disclosed herein improve functioning of computing systems that process audio input for speech recognition systems/engines. One or more aspects of the present disclosure propose systems and/or methods implementing a real-time blind source separation technique to provide improved automatic speech recognition based on concurrent multi-path processing of

audio signals. The proposed systems and method may achieve improved recognition accuracy even with the existence of one or a combination of interference, real-time algorithm delay, or a relatively small number of microphones (*e.g.*, 2 or 3). In particular, one or more aspects of the present disclosure provide improvements for distinguishing human utterances (e.g., speech) from noise or other sounds.

[0026] Blind source separation (BSS) is a technique for separating specific sources from sound mixtures without any information about the environment and sources. Depending on number of sources and microphones, BSS algorithms can be categorized into determined and underdetermined BSS. In a determined situation, the number of microphones is larger than or equal to the number of sources, and independent component analysis (ICA) is a commonly used method; in an underdetermined situation, where the number of microphones is less than the number of sources, nonnegative matrix factorization (NMF) has received much attention. One or more aspects of the disclosure presented herein propose a real-time blind source separation system to improve the ASR engine performance. The proposed solutions achieve good recognition accuracy with the existence of interference, with real-time algorithm delay, and small number of microphones.

[0027] FIG. 1 illustrates an example system 100 for concurrent multi-path processing of audio signals for automatic speech recognition, in accordance with various embodiments. The example system 100 may include a computing system 102 and an automatic speech recognition engine 112. The computing system 102 may include other components. The computing system 102 may include one or more processors and memory (*e.g.*, permanent memory, temporary memory). The processor(s) may be configured to perform various operations by interpreting machine-readable instructions stored in the memory. The computing system 102 may include other computing resources. The

computing system 102 may have access (*e.g.*, via one or more connections, via one or more networks) to other computing resources or other entities participating in the system 100.

[0028] The computing system 102 may include an input component 104, a first processing component 106, a second processing component 108, and an output component 110. The computing system 102 may include other components. While the computing system 102 is shown in FIG. 1 as a single entity, this is merely for ease of reference and is not meant to be limiting. One or more components or one or more functionalities of the computing system 102 described herein may be implemented in a single computing device or multiple computing devices. In some embodiments, one or more components or one or more functionalities of the computing system 102 described herein may be implemented in one or more networks, one or more endpoints, one or more servers, or one or more clouds.

[0029] The input component 104 may obtain audio information defining a set of audio signals. A set of audio signals may include multiple audio signals. The individual audio signals may be received from individual audio input devices. For example, individual audio signals may be received from individual audio input devices in a set of audio input devices included in a vehicle. An audio input device may include, for example, a microphone. Individual audio signals in the set of audio signals may convey mixed audio content produced by multiple audio sources. The audio sources may include one or a combination of humans, an ambient environment, or other sources. The audio content may include one or a combination of human utterances (*e.g.*, one or a combination of speech, words, sounds, or other utterances), background noise (*e.g.*, one or a combination of car noises, environment noises, or other noise), or other audio content. The human utterances may include commands intended for the ASR engine 112. The mixed audio content obtained by individual audio input

devices may refer to a combination of human utterances, background noise, and/or other sounds. The ASR engine 112 may be part of a device or other entity in which speech commands may be given. For example, the ASR engine 112 may control vehicle components based on commands uttered by a user within a vehicle. Accordingly, it may be desired to distinguish between individual human utterances and the background noise.

[0030] The output component 110 may determine a set of source-specific audio signals by demixing the mixed audio content produced by the multiple audio sources. Individual source-specific audio signals may represent individual audio content produced by specific individual audio sources of the multiple audio sources. For example, the demixing may result in distinguishing between utterances by different users and/or between human utterances and background noise. Determining the set of source-specific audio signals may be facilitated by processes and/or operations executed by one or both of the first processing component 106 and/or second processing component 108, described herein.

[0031] The first processing component 106 may be provided (*e.g.,* via input component 104) with the set of audio signals to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources. The processes performed by the first processing component 106 may be referred to herein as the "first signal processing path." The value of the demixing parameter may be in the form of a demixing matrix which may be initially estimated using independent vector analysis (IVA) and/or other techniques (see, *e.g.,* decomposition component 118).

[0032] The second processing component 108 may be provided (*e.g.,* via input component 104) with the set of audio signals to apply the value of the demixing parameter to the individual audio signals of the set of audio signals. Demixing may result in distinguishing, within individual audio signals, human utterance

from background noise. In some embodiments, the set of audio signals may be provided to the first processing component 106 and the second processing component 108 concurrently. The processes performed by the second processing component 108 may be referred to herein as the "second signal processing path."

[0033] FIG. 2 illustrates components utilized by first processing component 106 in order to achieve the determination of the value of the demixing parameter. The first processing component 106 may include of a transformation component 114, a collection component 115, a dimensionality component 116, a decomposition component 118, and a permutation component 120.

[0034] The transformation component 114 may be configured to transform, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals. In some implementations, the operation of transforming the individual audio signals into consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation and/or other operation. By way of non-limiting illustration, the individual audio signals in the set of audio signals may be passed through the analysis filter bank (AFB) to yield time-frequency (T-F) spectrum.

[0035] The collection component 115 may be configured to collect the consecutive frames of the time-frequency domain representation of the individual audio signals (*e.g.*, spectrum frames) into individual sets of combined consecutive frames of the time-frequency domain representations. By way of non-limiting illustration, the collection component 115 may utilize one or more buffers to collect the consecutive frames of the time-frequency domain representation of the individual audio signals.

[0036] The operations of the dimensionality component 116 and the decomposition component 118 described herein may be performed for

individual ones of the individual sets of combined consecutive frames of the time-frequency domain representations provided by the collection component 115. That is, once the consecutive frames of the time-frequency domain representation of the individual audio signals are collected into sets, the sets may be individually passed through dimensionality component 116 and the decomposition component 118.

[0037] The dimensionality component 116 may be configured to approximate individual reduced dimensionality representations of the individual sets of combined consecutive frames of the time-frequency domain representation of the individual audio signals. In some implementations, the operation of approximating individual reduced dimensionality representations may comprise a principal component analysis (PCA) whitening operation.

[0038] The source, observed, and separated signals in each time-frequency slot are described as $S_{k,t}$, $X_{k,t}$, and $Y_{k,t}$, respectively:

$$S_{k,t} = \left(S_{k,t,1}, \ldots, S_{k,t,N}\right)';$$

$$X_{k,t} = \left(X_{k,t,1}, \ldots, X_{k,t,M}\right)';$$

$$Y_{k,t} = \left(Y_{k,t,1}, \ldots, Y_{k,t,N}\right)';$$

where k and t are the frequency and frame indices, N and M are the number of sources and microphones. " $'$ "is the vector transpose, and the entries of these vectors are complex values.

[0039] As an illustrative example, consider the mixing and demixing of source signals in a 2-source and 2-microphone scenario. Each source may propagate different acoustic paths and arrives at the microphones. "A" represents the mixing matrix.

$$X_{k,t} = A_k S_{k,t}$$

[0040] Multichannel microphone signals X are given to a blind source separation (BSS) block and a demixing matrix W is estimated, and separated signals Y are obtained correspondingly.

[0041] Principal component analysis (PCA) whitening may comprise a pre-processing step. The goal of whitening may be to make the input less redundant. Given data X_m (*e.g.*, the individual sets of combined consecutive frames of the time-frequency domain representation of the individual audio signals), the covariance matrix may be computed as:

$$\mathbb{C}_{k,m1,m2} = \frac{1}{R}\sum_{t=1}^{R} X_{k,t,m1} X'_{k,t,m2}, \quad m1, m2 = 1, \ldots M;$$

where R is the number of frames. The principal direction of data variation, u_1, is the top eigenvector of $\mathbb{C}$, and similarly u_2 is the second eigenvector. Then the data may be rotated to maximize the independence as:

$$P = U'X = \begin{bmatrix} u_1'X \\ u_2'X \end{bmatrix}$$

[0042] The moving average (MA) may be applied to smooth the eigenvector, in order to remove the unnecessary disturbance on projection directions:

$$\overline{U} = \alpha\overline{U} + (1-\alpha)U;$$

where $\overline{U}$ and U are the smoothed and instantaneous eigenvectors.

[0043] The decomposition component 118 may be configured to decompose the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations and determine a current value of the demixing parameter based on the decomposed individual reduced dimensionality representations. In some implementations, the operations of decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations and determining the current value of the demixing parameter based on the decomposed individual reduced dimensionality representations may be facilitated by a multi-channel nonnegative matrix factorization.

[0044] Multi-channel nonnegative matrix factorization may utilize one or more of independent vector analysis (IVA), nonnegative matrix factorization (NMF), and/or other techniques.

[0045] Independent vector analysis (IVA) is a multivariate extension of independent component analysis (ICA) and can solve the permutation problem. The ICA based methods may only be applied to the determined situation. IVA assumes independence between the sources to estimate the demixing matrix, *'W'*. In addition, IVA assumes a spherical multivariate distribution as the source model to ensure higher-order correlations between frequency bins in each source.

[0046] Nonnegative matrix factorization (NMF) is a type of sparse representation algorithm that decomposes a nonnegative matrix into two nonnegative matrices as

$$D_{I\times J} = T_{I\times B}V_{B\times J};$$

where D is a nonnegative data matrix with dimension $I \times J$, T and V are basis matrix and activation matrix with dimension $I \times B$ and $B \times J$, respectively, I, J and B are number of frequency bins, number of frames, and number of bases when NMF is applied to an acoustic signal.

[0047] Demixing matrix, W, may be updated based on the following rules:

$$D_{k,m} = \frac{1}{R}\sum_{t=1}^{R}\frac{1}{r_{k,t,m}}X_{k,t}X_{k,t}^{H};$$

$$W_{k,m} = \left(W_k D_{k,m}\right)^{-1} e_m;$$

$$W_{k,m} = W_{k,m}\left(W_{k,m}^{H} D_{k,m} W_{k,m}\right)^{-\frac{1}{2}}; \text{ and}$$

$$Q = WP.$$

[0048] NMF basis matrix T and activation matrix V, may be updated by following the Itakura-Saito divergence update rules:

$$T = T\sqrt{\frac{((TV)^{-2}|Q|^2)V'}{(TV)^{-1}V'}};$$

$$V = V\sqrt{\frac{T'((TV)^{-2}|Q|^2)}{T'(TV)^{-1}}}; \text{ and}$$

$$r_{k,t,m} = \sum_b T_{k,b,m}V_{b,t,m}.$$

[0049] The permutation component 120 may be configured to compare the current value of the demixing parameter to a previous value of the demixing parameter; and based on the comparison, set the value of the demixing parameter as either the current value or a modified value of the current value (e.g., an aligned demixing matrix, described below). The operations of permutation component 120 may be directed to solving a permutation problem. Permutation may refer to an unmixing of mixed audio signals into separate output channels (where an individual output channel includes either a speech or a noise audio signal) where an output channel previously determined to include the speech audio signal may subsequently be determined to be a noise signal. For example, for a given frame, it may be determined that a first output channel includes an unmixed audio signal comprising speech, and a second output channel includes an unmixed audio signal comprising noise. Permutation may occur when, for a subsequent frame, it may be determined that the first output channel includes an unmixed audio signal comprising noise, and the second output channel includes an unmixed audio signal comprising speech. Thus, the output channels may be considered permutated.

[0050] In previous solutions, it may be assumed that mixing and demixing matrix are time invariant, which often does not hold in practice. The permutation component 120 may track this change with a certain delay, that is, buffer frame size. Adjusting buffer frame size is a trade-off between stability and time sensitivity. Larger buffer frame size may guarantee better performance but less time sensitivity, and vice versa. Permutation may occur

when buffer size is too small. For example, in a short period only one source is active, and demixing matrix which is updated on data from one source can be easily biased, or even worse, permutated.

[0051] Two approaches are proposed to solve the permutation issue. In a spatial approach, comparing the current value of the demixing parameter to a previous value of the demixing parameter may refer to calculating a distance between a current demixing matrix (W) and a previous demixing matrix. If different sources separate apart sufficiently, the distance matrix may be close to diagonal. If the distance matrix is far away from diagonal, there is high probability that only one source is active, and then permutation component 120 proceeds to a statistical approach for further examination.

[0052] For the statistic approach, the basis matrix in NMF represents the frequently appearing spectral patterns. The basis matrix can be utilized to separate speech from interferences, *e.g.*, background noise, music, etc. By classifying the basis matrix, one can decide whether the active source is from desired speech or interference (e.g., noise, non-speech, etc.) and align the demixing matrix accordingly. Classifying the basis matrix may include determining whether the basis matrix is indicative of speech or noise. This may be accomplished by comparing a given basis matrix to a basis matrix known to be indicative of speech. For example, speech within the time-frequency domain may be more concentrated in low frequencies, display a harmonic structure, and/or have low energy. Noise within the time-frequency domain may be flat and/or have substantially even power in low and high frequencies. Aligning the demixing matrix may produce the modified value of the demixing parameter, *e.g.*, the aligned demixing matrix. Aligning the basis matrix may correct permutation problem by changing output channels to conform to one standard. For example, as described above permutation may occur where output is mixed,

and aligning the demixing matrix may reposition entries in the matrix to ensure the output channels are consistent (e.g., not permutated).

[0053] FIG. 3 illustrates components utilized by second processing component 108 in order to achieve the application of the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path. The second processing component 108 may include one or a combination of a transformation component 122, a dimensionality component 124, a demixing component 126, or other components.

[0054] The transformation component 122 may be configured to transform, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals. The operation of transforming the individual audio signals to consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

[0055] The operations of the dimensionality component 124 and the demixing component 126 described herein may be performed consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals. This process differs from the first signal processing path which collected the consecutive frames into combined sets, thus causing some delay in the processing to finish the collection.

[0056] The dimensionality component 124 may be configured to approximate, for the individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, an individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signals included in an individual frame. The operation of approximating the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the

individual frame comprises a principal component analysis (PCA) whitening operation, described herein.

[0057] The demixing component 126 may be configured to apply, for the individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, the value of the demixing parameter (obtained from the first processing component 106) to the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signals included in the individual frames. The application of the value of the demixing parameter may be accomplished through matrix multiplication. The application of the value of the demixing parameter may obtain, for the individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, an individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signals included in the individual frames. The demixing component 126 may pass its output to the output component 110.

[0058] FIG. 4 illustrates components utilized by output component 110 in order to achieve the generation of the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals. The output component 110 may include one or a combination of a dimensionality component 128, a transformation component 130, a transmission component 132, or other components.

[0059] The operations of the dimensionality component 128 and the transformation component 130 described herein may be performed consecutively for individual frames in the consecutive frames of individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signals included in the individual frames.

[0060] The dimensionality component 128 may be configured to restore, for the individual frames, the dimensionality of the individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame. Restoring the dimensionality may obtain individual demixed time-frequency domain representation of the individual audio signals included in the individual frames. This processes of dimensionality component 128 may comprise a normalization that restores the signal scale back to the original power by applying a back-projection technique and/or other technique.

[0061] The transformation component 130 may be configured to transform the individual demixed time-frequency domain representation of the individual audio signals included in the individual frames to a time domain representation. The time domain representation of the consecutive frames of the individual demixed time-frequency domain representation of the individual audio signals may generate the individual source-specific audio signals. The operation of transforming to the time domain representation may comprise a synthesis filter bank (SFB) operation. By way of non-limiting illustration, the individual demixed time-frequency domain representation of the individual audio signals included in the individual frames may be passed through SFB to restore the signal to time domain and generate the individual source-specific audio signals (*e.g.*, the signals have been demixed, restored to the original power, and transformed back into time domain).

[0062] The transmission component 132 may be configured to transmit the individual source-specific audio signals to the ASR engine 112. A source-specific audio signal may be transmitted directly or indirectly to the ASR engine 112 by the transmission component 132.

[0063] FIG. 5 illustrates an example flow chart 200 for concurrent multi-path processing of audio signals for automatic speech recognition, in accordance

with various embodiments of the disclosure. At block 502, audio information defining a set of audio signals may be obtained. The individual audio signals in the set of audio signals may convey mixed audio content produced by multiple audio source. At block 504, a set of source specific audio signals may be obtained by demixing the mixed audio content produced by the multiple audio sources. Individual source-specific audio signals may represent individual audio content produced by specific individual audio sources of the multiple audio sources. Blocks 506-510 illustrate the processes for determining the set of source-specific audio signals. At block 506, the set of audio signals to may be provided to a first signal processing path to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources. At block 508, the set of audio signals may be concurrently provided to a second signal processing path to apply the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path. At block 510 the individual source-specific audio signals may be generated from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals.

[0064] FIG. 6 illustrates an example flow diagram of concurrent multi-path processing of audio signals for automatic speech recognition, in accordance with various embodiments of the disclosure. In particular, the FIG. 6 illustrates and distinguishes between the first signal processing path and the second signal processing path presented herein. Elements 601 and 602 represent two audio input devices (microphones). Element 601 generates an audio signal conveying mixed audio content produced by multiple audio sources. Element 602 generates an audio signal conveying mixed audio content produced by multiple audio sources. A goal of the signal processing is to determine source-specific audio signals representing audio content produced by individual ones of the

sources. For illustrative purposes, we will assume the audio signals generated by elements 601 and 601 are representative of audio mixed by two sources. Thus, a goal of the processing may be to obtain a first source-specific audio signal 628 representative of audio content generated by one source, and a second source specific audio signal 630 representative of audio content generated by the other source.

[0065] The audio signals from elements 601 and 602 may be provided to analysis filter bank (AFB) 604 and AFB 606, respectively, to yield time-frequency (T-F) spectrum (*e.g.*, consecutive frames of the time-frequency domain representation of the individual audio signals). The processing through elements 608 – 616 may represent the first signal processing path. The processing through elements 618-620 may represent the second signal processing path. The processing through elements 622-626 may represent processes carried out by output component 110 (shown in FIG. 1 and described herein).

[0066] Referring to the first signal processing path, elements 608 and 610 may represent individual buffers configured to collect the frequency domain representations after passing through elements 604 and 606, respectively. These buffers may represent the operation of collecting the consecutive frames of the time-frequency domain representation of the individual audio signals into individual sets of combined consecutive frames of the time-frequency domain representation.

[0067] Element 612 may represent a principal component analysis (PCA) whitening operation and/or other operation configured to approximate individual reduced dimensionality representations of the combined consecutive frames of the time-frequency domain representation of the individual audio signals.

[0068] Element 614 may perform tasks such as decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations and determining a current value of the demixing parameter based on the decomposed individual reduced dimensionality representations. By way of non-limiting illustration, element 614 may represent a multi-channel nonnegative matrix factorization.

[0069] Element 616 may serve to solve the permutation problem. By way of non-limiting illustration, element 616 may perform one or more of comparing the current value of the demixing parameter to a previous value of the demixing parameter; and based on the comparison, setting the value of the demixing parameter as either the current value or a modified version of the current value.

[0070] Element 618 may obtain consecutive frames of the time-frequency domain representation of the individual audio signals after passing through AFBs 604 and 606. The element 618 may consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, approximate an individual reduced dimensionality representation of the time-frequency domain representation of an individual audio signal included in an individual frame. Element 618 may represent a principal component analysis (PCA) whitening operation within the second signal processing path. Line 605 may represent the operations:

$$P = U'X = \begin{bmatrix} u_1'X \\ u_2'X \end{bmatrix}; \text{ and}$$

$$\overline{U} = \alpha\overline{U} + (1-\alpha)U,$$

described above.

[0071] Element 620 may consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, apply the value of the demixing parameter (demixing matrix obtain from element 616) to the individual reduced dimensionality

representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame.

[0072] Element 622 may consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals obtain from element 620, restore the dimensionality of the individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed time-frequency domain representation of the individual audio signal included in the individual frame. By way of non-limiting illustration, element 622 may represent a back-projection technique and/or other technique to restore the signals to original dimensionality.

[0073] Elements 624 and 626 may consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals, transform the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to a time domain representation. The time domain representation of the consecutive frames of the individual demixed time-frequency domain representation of the individual audio signals may the first source-specific audio signal 628 and the second source-specific audio signal 630. By way of non-limiting illustration, elements 624 and 626 may represent synthesis filter bank operations. It is noted that although elements 604, 606, 624, and 626 are shown as separate elements within FIG. 6, this is for illustrative purposes only. In some implementations, element 604 (analysis filter bank) and element 624 (synthesis filter bank) may be integrated into a single component and element

606 (analysis filter bank) and element 626 (synthesis filter bank) may be integrated into a single component.

[0074] Experimentation has shown improvement over prior techniques with regard to delay and word error rate (WER) in speech processing when audio is obtained from both a back seat and a front seat of a vehicle. A higher WER represents worse performance. For example, with raw input (*e.g.*, without signal separation) WER is highest. For a known prior technique of signal separation (*See, e.g.*, D. Kitamura, *et al*. 'Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,' IEEE Trans. ASLP, vol. 24, no. 9, pp. 1626-1641, 2016) WER is improved by about half that of using raw input with respect to both front and back seat, and with delay of multiple seconds. With the solution proposed by the present disclosure, WER is improved by about two to six words in the back seat, and about one to two words for the front seat. Further, delay has been shown to be only in the thousandths of seconds. Both WER and delay are significantly improved.

[0075] FIG. 7 is a block diagram that illustrates a computer system 700 upon which any of the embodiments described herein may be implemented. The computer system 700 includes a bus 702 or other communication mechanism for communicating information, one or more hardware processors 704 coupled with bus 702 for processing information. Hardware processor(s) 704 may be, for example, one or more general purpose microprocessors.

[0076] The computer system 700 also includes a main memory 706, such as a random access memory (RAM), cache and/or other dynamic storage devices, coupled to bus 702 for storing information and instructions to be executed by processor(s) 704. Main memory 706 also may be used for storing temporary variables or other intermediate information during execution of instructions to be executed by processor(s) 704. Such instructions, when stored in storage

media accessible to processor(s) 704, render computer system 700 into a special-purpose machine that is customized to perform the operations specified in the instructions. Main memory 706 may include non-volatile media and/or volatile media. Non-volatile media may include, for example, optical or magnetic disks. Volatile media may include dynamic memory. Common forms of media may include, for example, a floppy disk, a flexible disk, hard disk, solid state drive, magnetic tape, or any other magnetic data storage medium, a CD-ROM, any other optical data storage medium, any physical medium with patterns of holes, a RAM, a DRAM, a PROM, and EPROM, a FLASH-EPROM, NVRAM, any other memory chip or cartridge, and networked versions of the same.

[0077] The computer system 700 may implement the techniques described herein using customized hard-wired logic, one or more ASICs or FPGAs, firmware and/or program logic which in combination with the computer system causes or programs computer system 700 to be a special-purpose machine. According to one embodiment, the techniques herein are performed by computer system 700 in response to processor(s) 704 executing one or more sequences of one or more instructions contained in main memory 706. Such instructions may be read into main memory 706 from another storage medium, such as storage device 708. Execution of the sequences of instructions contained in main memory 706 causes processor(s) 704 to perform the process steps described herein. For example, the process/method shown in FIG. 5 and/or FIG. 6 and described in connection with these figures can be implemented by computer program instructions stored in main memory 706. When these instructions are executed by processor(s) 704, they may perform the steps as shown in FIG. 5 and/or FIG. 6 and described above. In alternative embodiments, hard-wired circuitry may be used in place of or in combination with software instructions.

[0078] The computer system 700 also includes a communication interface 710 coupled to bus 702. Communication interface 710 provides a two-way data communication coupling to one or more network links that are connected to one or more networks. As another example, communication interface 710 may be a local area network (LAN) card to provide a data communication connection to a compatible LAN (or WAN component to communicated with a WAN). Wireless links may also be implemented.

[0079] The performance of certain of the operations may be distributed among the processors, not only residing within a single machine, but deployed across a number of machines. In some example embodiments, the processors or processor-implemented engines may be located in a single geographic location (*e.g.*, within a home environment, an office environment, or a server farm). In other example embodiments, the processors or processor-implemented engines may be distributed across a number of geographic locations.

[0080] Certain embodiments are described herein as including logic or a number of components. Components may constitute either software components (*e.g.*, code embodied on a machine-readable medium) or hardware components (*e.g.*, a tangible unit capable of performing certain operations which may be configured or arranged in a certain physical manner). As used herein, for convenience, components of the computing system 102 may be described as performing or configured for performing an operation, when the components may comprise instructions which may program or configure the computing system 102 to perform the operation.

[0081] While examples and features of disclosed principles are described herein, modifications, adaptations, and other implementations are possible without departing from the spirit and scope of the disclosed embodiments. Also, the words "comprising," "having," "containing," and "including," and other similar forms are intended to be equivalent in meaning and be open ended in

that an item or items following any one of these words is not meant to be an exhaustive listing of such item or items, or meant to be limited to only the listed item or items. It must also be noted that as used herein and in the appended claims, the singular forms "a," "an," and "the" include plural references unless the context clearly dictates otherwise.

[0082] The embodiments illustrated herein are described in sufficient detail to enable those skilled in the art to practice the teachings disclosed. Other embodiments may be used and derived therefrom, such that structural and logical substitutions and changes may be made without departing from the scope of this disclosure. The Detailed Description, therefore, is not to be taken in a limiting sense, and the scope of various embodiments is defined only by the appended claims, along with the full range of equivalents to which such claims are entitled.

WHAT IS CLAIMED IS:

1. A system for concurrent multi-path processing of audio signals for automatic speech recognition, the system comprising:

one or more processors; and

a memory storing instructions that, when executed by the one or more processors, cause the system to perform:

obtaining audio information defining a set of audio signals, individual audio signals in the set of audio signals conveying mixed audio content produced by multiple audio sources; and

determining a set of source-specific audio signals by demixing the mixed audio content produced by the multiple audio sources, individual source-specific audio signals representing individual audio content produced by specific individual audio sources of the multiple audio sources, wherein determining the set of source-specific audio signals comprises:

providing the set of audio signals to a first signal processing path to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources;

concurrently providing the set of audio signals to a second signal processing path to apply the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path; and

generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals.

2. The system of claim 1, wherein providing the set of audio signals to the first signal processing path further includes the operations of:

transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals;

collecting the consecutive frames of the time-frequency domain representation of the individual audio signals into individual sets of combined consecutive frames of the time-frequency domain representations; and

for the individual sets of combined consecutive frames of the time-frequency domain representations:

approximating individual reduced dimensionality representations of the individual sets of combined consecutive frames of the time-frequency domain representation of the individual audio signals;

decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations;

determining a current value of the demixing parameter based on the decomposed individual reduced dimensionality representations;

comparing the current value of the demixing parameter to a previous value of the demixing parameter; and

based on the comparison, setting the value of the demixing parameter as either the current value or a modified version of the current value.

3. The system of claim 2, wherein the operation of transforming the individual audio signals into consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

4. The system of claim 2, wherein the operation of approximating individual reduced dimensionality representations comprises a principal component analysis (PCA) whitening operation.

5. The system of claim 2, wherein the operation of decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations comprises a multi-channel nonnegative matrix factorization.

6. The system of claim 1, wherein providing the set of audio signals to the second signal processing path further includes the operations of:

transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals; and

consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals:

approximating an individual reduced dimensionality representation of the time-frequency domain representation of an individual audio signal included in an individual frame; and

applying the value of the demixing parameter to the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame.

7. The system of claim 6, wherein the operation of transforming the individual audio signals to consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

8. The system of claim 6, wherein the operation of approximating the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame comprises a principal component analysis (PCA) whitening operation.

9. The system of claim 6, wherein generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals comprises operations of:

consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals:

restoring the dimensionality of the individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed time-frequency domain representation of the individual audio signal included in the individual frame; and

transforming the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to a time domain representation; and

wherein the time domain representation of the consecutive frames of the individual demixed time-frequency domain representation of the individual audio signals comprises the individual source-specific audio signals.

10. The system of claim 9, wherein the operation of transforming the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to the time domain representation comprises a synthesis filter bank (SFB) operation.

11. A method for concurrent multi-path processing of audio signals for automatic speech recognition, the method comprising:

obtaining audio information defining a set of audio signals, individual audio signals in the set of audio signals conveying mixed audio content produced by multiple audio sources; and

determining a set of source-specific audio signals by demixing the mixed audio content produced by the multiple audio sources, individual source-specific audio signals representing individual audio content produced by specific individual audio sources of the multiple audio sources, wherein determining the set of source-specific audio signals comprises:

providing the set of audio signals to a first signal processing path to determine a value of a demixing parameter for demixing the mixed audio content produced by the multiple audio sources;

concurrently providing the set of audio signals to a second signal processing path to apply the value of the demixing parameter to the individual audio signals of the set of audio signals provided to the second signal processing path; and

generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals.

12. The method of claim 11, wherein providing the set of audio signals to the first signal processing path further includes operations of:

transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals;

collecting the consecutive frames of the time-frequency domain representation of the individual audio signals into individual sets of combined consecutive frames of the time-frequency domain representations; and

for the individual sets of combined consecutive frames of the time-frequency domain representations:

approximating individual reduced dimensionality representations of the individual sets of combined consecutive frames of the time-frequency domain representation of the individual audio signals;

decomposing the individual reduced dimensionality representations into individual sets of spectral bases and temporal activations;

determining a current value of the demixing parameter based on the decomposed individual reduced dimensionality representations;

comparing the current value of the demixing parameter to a previous value of the demixing parameter; and

based on the comparison, setting the value of the demixing parameter as either the current value or a modified version of the current value.

13. The method of claim 12, wherein the operation of transforming the individual audio signals into consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

14. The method of claim 12, wherein the operation of approximating individual reduced dimensionality representations comprises a principal component analysis (PCA) whitening operation.

15. The method of claim 12, wherein the operation of decomposing the individual reduced dimensionality representations into individual sets of

spectral bases and temporal activations comprises a multi-channel nonnegative matrix factorization.

16. The method of claim 11, wherein the second signal processing path includes operations of:

transforming, based on a transformation rate, the individual audio signals to consecutive frames of a time-frequency domain representation of the individual audio signals; and

consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals:

approximating an individual reduced dimensionality representation of the time-frequency domain representation of an individual audio signal included in an individual frame; and

applying the value of the demixing parameter to the individual reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame.

17. The method of claim 16, wherein the operation of transforming the individual audio signals to consecutive frames of the time-frequency domain representation of the individual audio signals comprises an analysis filter bank (AFB) operation.

18. The method of claim 16, wherein the operation of approximating the individual reduced dimensionality representation of the time-frequency domain

representation of the individual audio signal included in the individual frame comprises a principal component analysis (PCA) whitening operation.

19. The method of claim 16, wherein generating the individual source-specific audio signals from the individual audio signals based on the application of the value of the demixing parameter to the individual audio signals comprises operations of:

consecutively, for individual frames in the consecutive frames of the time-frequency domain representation of the individual audio signals:

restoring the dimensionality of the individual demixed reduced dimensionality representation of the time-frequency domain representation of the individual audio signal included in the individual frame to obtain an individual demixed time-frequency domain representation of the individual audio signal included in the individual frame; and

transforming the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to a time domain representation; and

wherein the time domain representation of the consecutive frames of the individual demixed time-frequency domain representation of the individual audio signals comprises the individual source-specific audio signals.

20. The method of claim 19, wherein the operation of transforming the individual demixed time-frequency domain representation of the individual audio signal included in the individual frame to the time domain representation comprises a synthesis filter bank (SFB) operation.

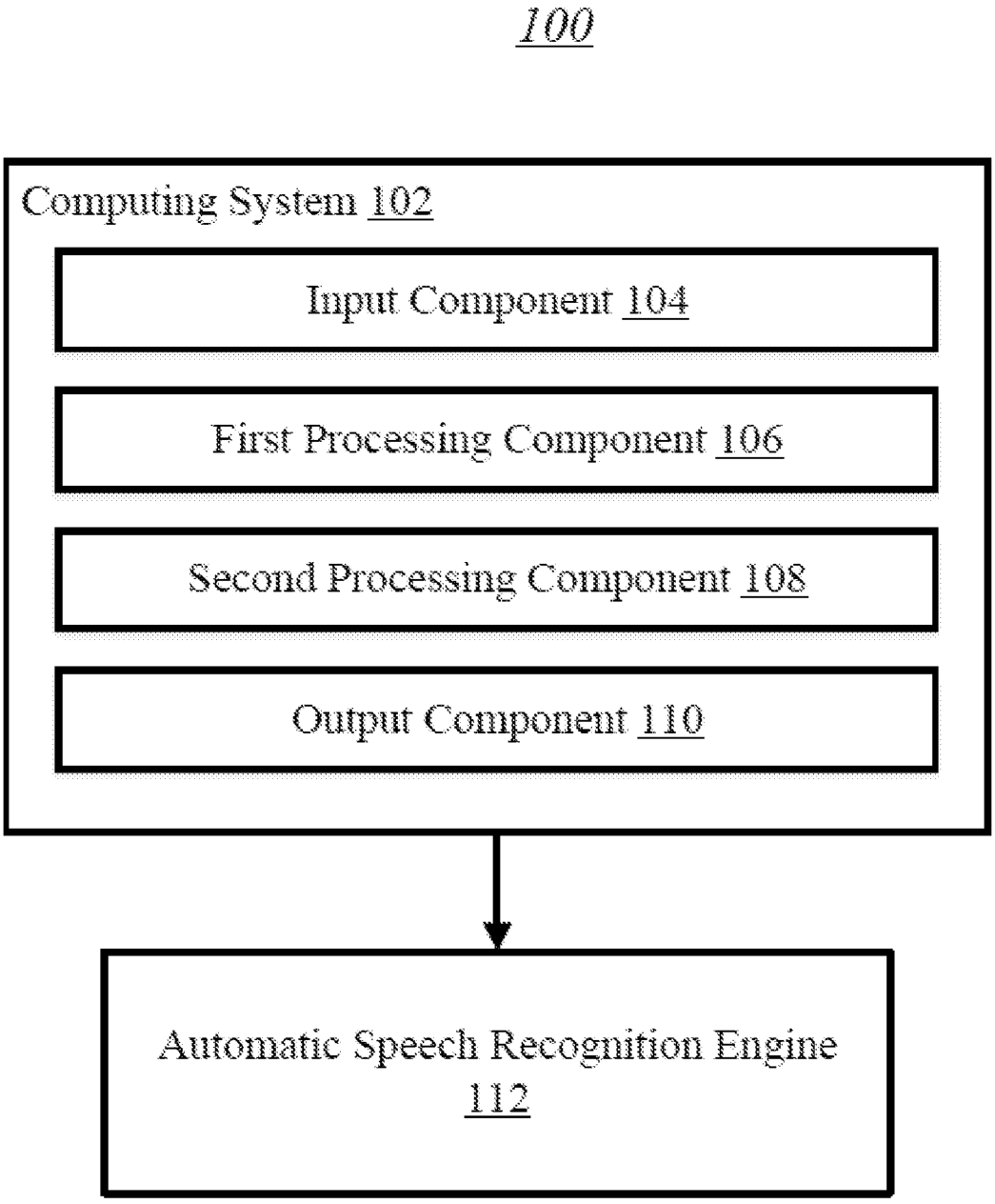
100
Computing System 102
Input Component 104
First Processing Component 106
Second Processing Component 108
Output Component 110
Automatic Speech Recognition Engine 112

FIG. 1

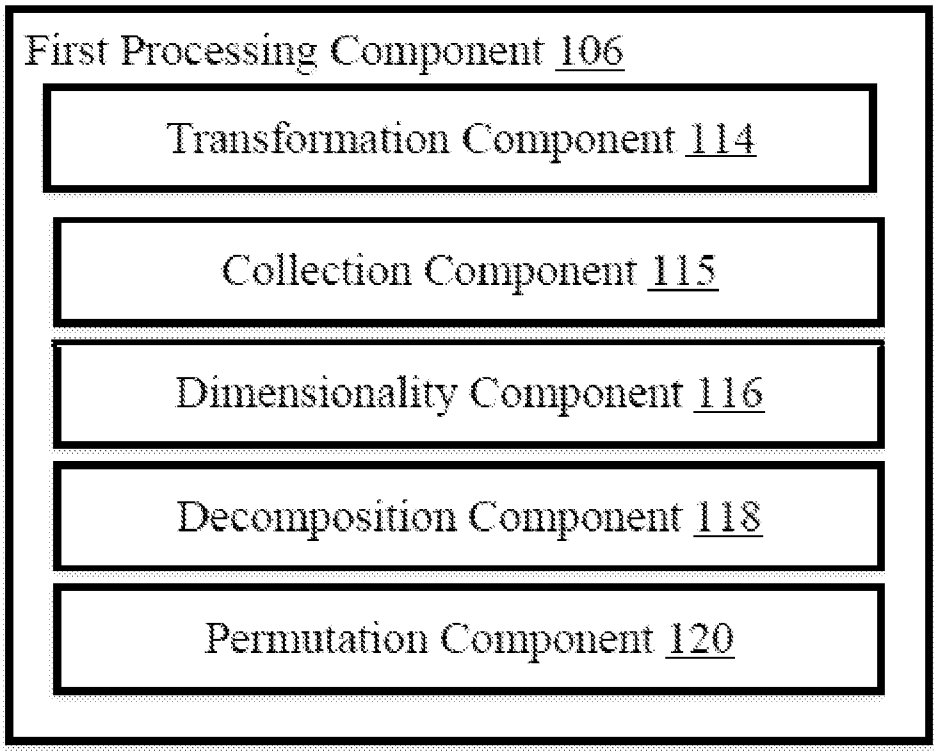
First Processing Component 106
Transformation Component 114
Collection Component 115
Dimensionality Component 116
Decomposition Component 118
Permutation Component 120

FIG. 2

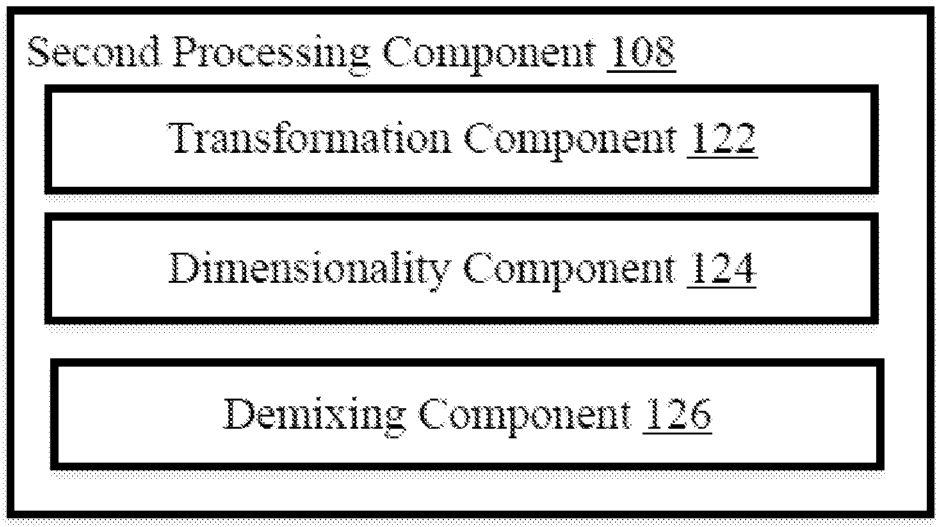
Second Processing Component 108
Transformation Component 122
Dimensionality Component 124
Demixing Component 126

FIG. 3

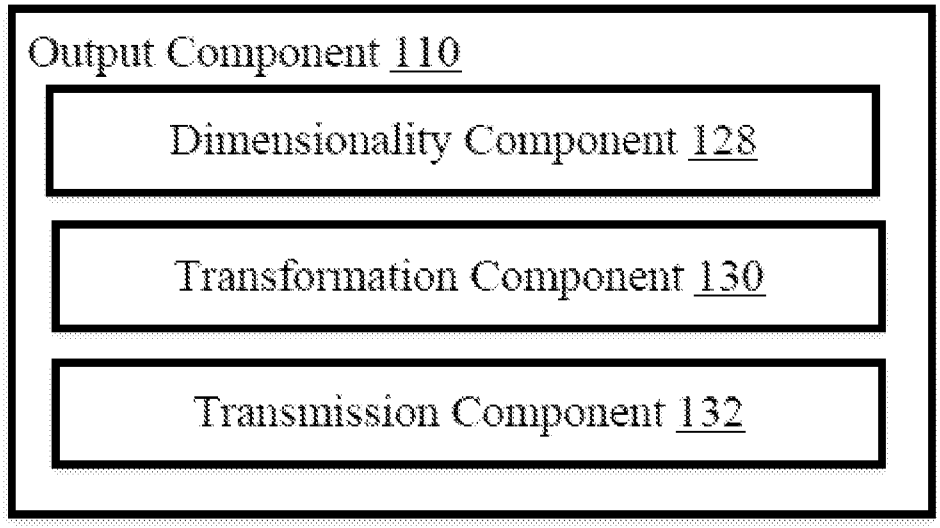
Output Component 110
Dimensionality Component 128
Transformation Component 130
Transmission Component 132

FIG. 4

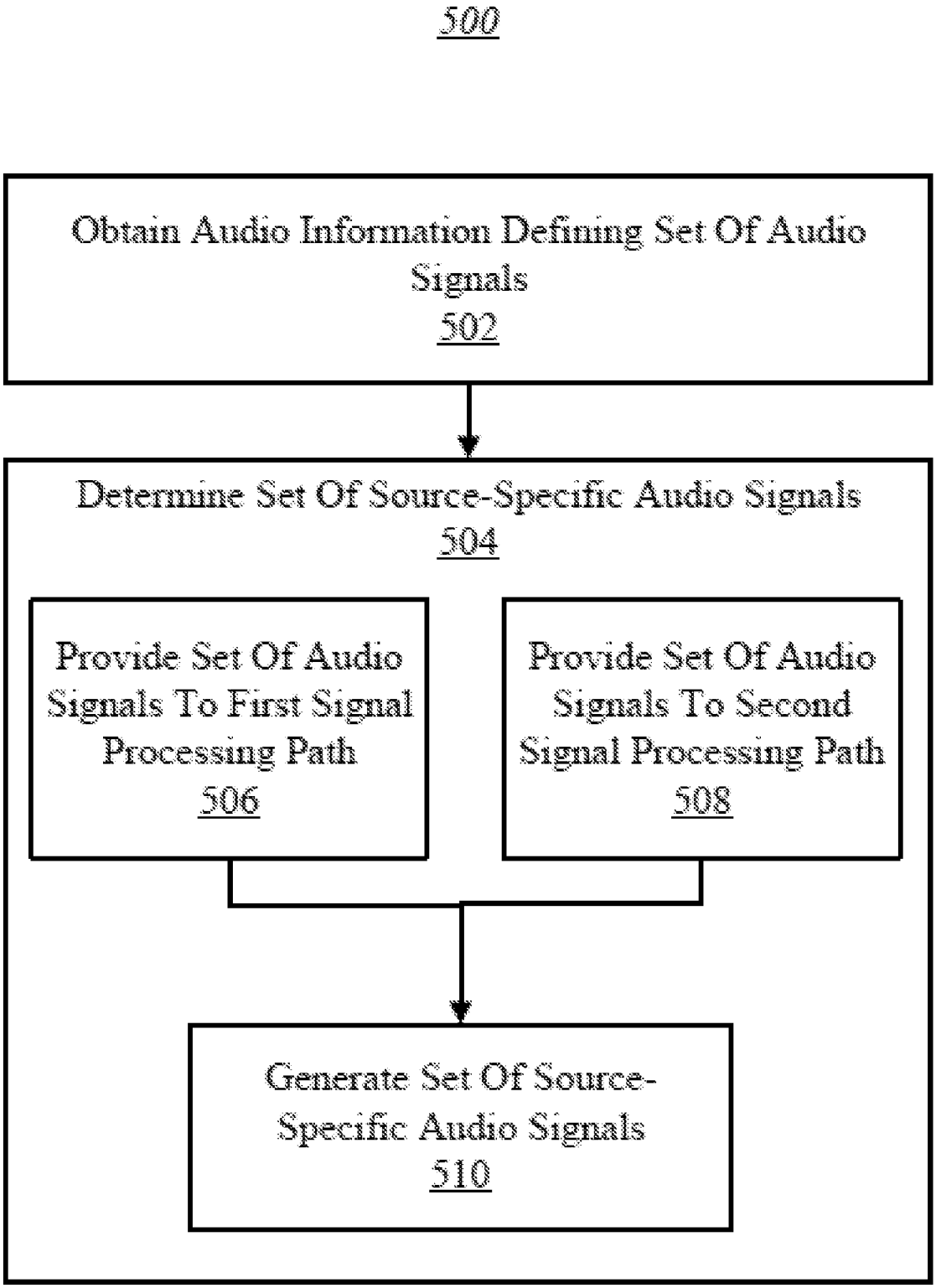
500
Obtain Audio Information Defining Set Of Audio Signals
502
Determine Set Of Source-Specific Audio Signals
504
Provide Set Of Audio Signals To First Signal Processing Path
506
Provide Set Of Audio Signals To Second Signal Processing Path
508
Generate Set Of Source-Specific Audio Signals
510

FIG. 5

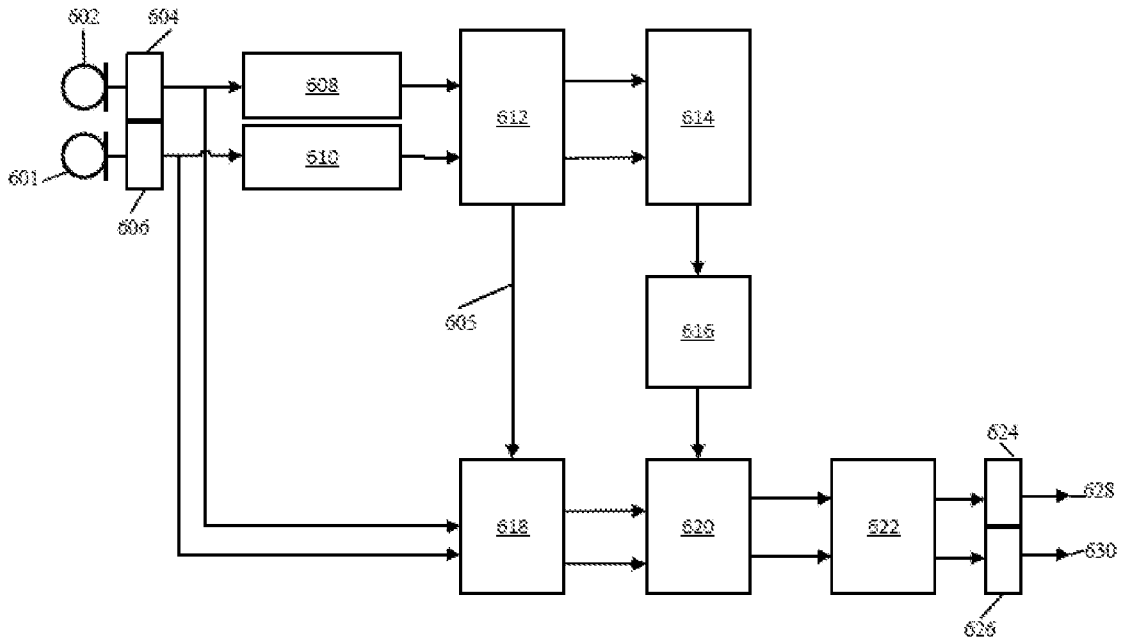

FIG. 6

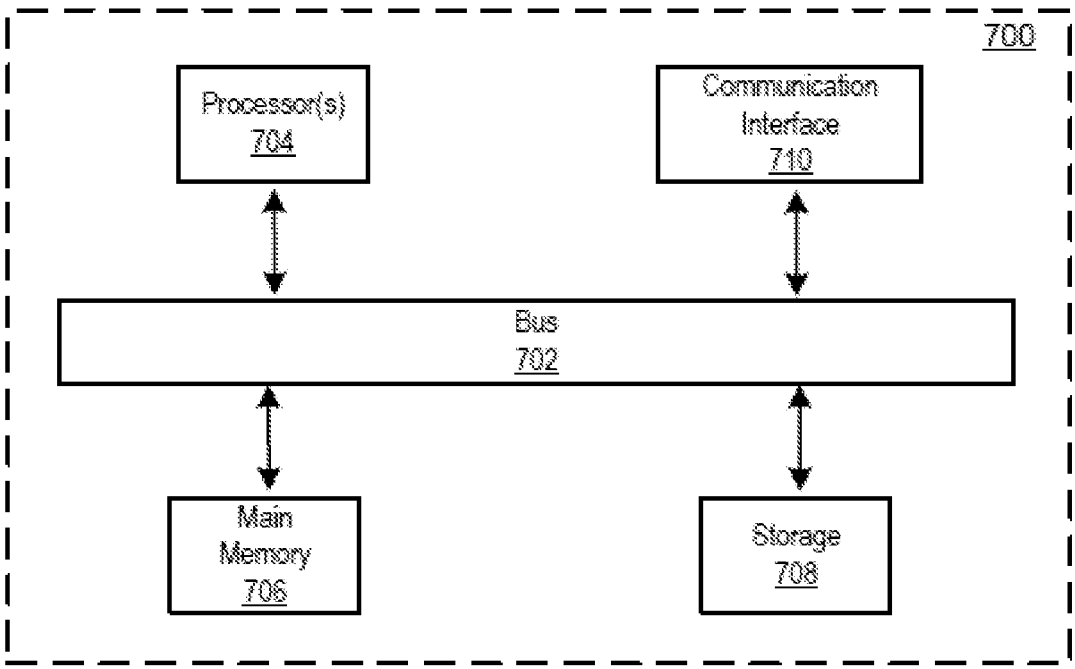
700
Processor(s)
704
Communication
Interface
710
Bus
702
Main
Memory
706
Storage
708

FIG. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/076395

A. CLASSIFICATION OF SUBJECT MATTER

G10L 21/0272(2013.01)i

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)

G10L 21; G10L 25

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CNKI, CNTXT, VEN: mix+, audio, sound, acoust+, vocal+, demix+, unmix+, de-mix+, un=mix+, separate+, multi+ source?, plural, councurrent+, paral+, matrix

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 9668066 B1 (CEDAR AUDIO LTD.) 30 May 2017 (2017-05-30) the description, column 2 line 1 to line 37, column 3 line 1 to line 46, column 4 line 55 to column 10 line 29, column 12 line 65 to column 15 line 56, column 19 line 40 to column 20 line 64, claims 1-20	1-20
X	CN 101996639 A (SPRING FOUND NCTU) 30 March 2011 (2011-03-30) the description, paragraphs [0032]-[0044]	1-20
X	JP 2005236852 A (NIPPON HOSO KYOKAI KK.) 02 September 2005 (2005-09-02) the description, paragraphs [0030]-[0076]	1-20
X	CN 1808571 A (MATSUSHITA ELECTRIC IND. CO., LTD.) 26 July 2006 (2006-07-26) the description, pages 7-10	1-20
A	US 2012045066 A1 (HONDA MOTOR CO., LTD.) 23 February 2012 (2012-02-23) the whole document	1-20
A	US 2007025564 A1 (KOBE SEIKO SHO KK.) 01 February 2007 (2007-02-01) the whole document	1-20

☐ 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
28 November 2019	**06 December 2019**

Name and mailing address of the ISA/CN	Authorized officer
National Intellectual Property Administration, PRC **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing** **100088** **China**	**YANG,Shilin**
Facsimile No. **(86-10)62019451**	Telephone No. **86-（010）-62085717**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/076395

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
US 9668066 B1	30 May 2017	None	
CN 101996639 A	30 March 2011	CN 101996639 B	06 June 2012
JP 2005236852 A	02 September 2005	None	
CN 1808571 A	26 July 2006	WO 2006078003 A2	27 July 2006
		WO 2006078003 A3	08 February 2007
US 2012045066 A1	23 February 2012	JP 5706782 B2	22 April 2015
		JP 2012042953 A	01 March 2012
		US 8867755 B2	21 October 2014
US 2007025564 A1	01 February 2007	EP 1748588 A2	31 January 2007
		EP 1748588 A3	27 February 2008
		JP 2007034184 A	08 February 2007