

DESCRIPTION

Title of the invention: Method and system for separating in real-time, in an audio stream, the voice component and the noise component

[0001] The invention relates to a method and system allowing, in an audio stream, the portion of the stream associated with a voice or speech to be separated in real-time from a portion containing noise.

[0002] The invention is applicable to a context in which one or more people are speaking in a noisy environment (hubbub, engine noise, fans, etc.). The speech signal superposed on noisy signals is digitized into an audio stream by an audio sensor.

[0003] The invention also relates to a method and a system for enhancing a real-time voice signal in an audio stream by using a method for separating audio sources in delayed time.

[0004] The prior art known to the applicant may be divided into two categories: approaches that may be said to be conventional, and approaches made possible by the artificial-intelligence technique known as deep learning.

[0005] With respect to approaches employing deep learning, some treat the problem of separating voice/background noise directly whereas others address signal/signal and voice/voice separation.

[0006] Patent application US 20190066713 discloses a method consisting in obtaining, with a device, a combined audio signal from combined signals originating from multiple audio sources in a region in which a person is found. The implemented processing makes use of neuronal processing.

[0007] An example of a method for separating a plurality of voices in an audio signal according to the prior art comprises the steps described below, which have not been illustrated for simplicity reasons. The input audio signal is denoted X ; it has a length L .

The signal is transmitted to an encoder M_1 that converts X into a tensor $X^{(1)}$ of $F \times T$ size, where T is a divisor of L and F a number of filters as set by the designer. The encoder M_1 is a 1D, F -filter convolution. The coefficients of the convolution kernels are adjusted in a training phase. The tensor is transmitted on the one hand to a multiplier for future use and on the other hand to a separating module. The separating module is divided into two submodules M_2 and M_4 . The first submodule M_2 converts the tensor $X^{(1)}$ into a tensor $X^{(2)}$ of $F \times T$ size. The first submodule M_2 is

made up of a normalisation layer, a 1x1 convolution and a stack of 1D-Conv modules that are known in the art and the parameters of which are adjusted in a training phase.

[0008] The second submodule M_4 converts $X^{(2)}$ into $X^{(4)}$, a tensor of $2F \times T$ size. To do this, the second submodule M_4 applies in sequence a nonlinearity, a 1x1 convolution and a sigmoid function. The coefficients of the 1x1 convolution are adjusted in a training phase.

[0009] $X^{(1)}$ is concatenated with itself to form a tensor of $2F \times T$ size that is multiplied by $X^{(4)}$ to form $X^{(5)}$.

[0010] The module M_5 receives $X^{(5)}$ as input and delivers as output two signals of L length by means of a 1D deconvolution the parameters of which are adjusted in a training phase.

[0011] The digital parameters defining the processing of the different modules are obtained in a prior learning phase on a database.

[0012] By replacing one of the voices with noise, it is immediate to use the methods described in the state of the art to separate the voice from the background noise in an audio signal and, retaining only the output containing the voice signal, to enhance the voice of a noisy signal.

[0013] Figure 1 illustrates an application to the separation of signals of various types, here the separation of a voice channel and a noise channel.

[0014] As described, the technical teaching of the prior art does not directly allow real-time processing of an audio stream.

[0015] The document by Mimilakis Stylianos Ioannis et al, entitled "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation", 25 September 2017, pages 1-6, XP 033263882, discloses a method allowing voice to be separated from a musical background.

[0016] In the field of deep-learning techniques, data are represented in the form of tensors. The data are modified by a succession of modules. At the output of each module, the data are projected into an abstract space that is in general defined by its dimensions. Between the modules, the manipulated intermediate data contain information on the input data.

[0017] To do this, the present invention implements the following treatments:

[0018] The signal (input stream X) is divided into N frames of length L , with X_N the n th frame. The method executes the following processing operations:

[0019] The frame X_N is encoded via a 1D convolution neural network. The result is a tensor $X_N^{(1)}$ of $F \times T$ size, with F the number of filters as given by the designer and T a divisor of L that is dependent on the size of the filters F , 100. The result $X_N^{(1)}$ is then converted by a module M_2 , 101. The result $X_N^{(2)}$ is a tensor of $F \times T$ size. The module M_4 estimates, 103, from $X_N^{(2)}$, a tensor $X_N^{(4)}$ of $2F \times T$ size. $X_N^{(1)}$ is concatenated with itself, 104, to form a tensor of $2F \times T$ size that is multiplied by $X_N^{(4)}$ to form $X_N^{(5)}$. The module M_5 produces from $X_N^{(5)}$ a tensor of $2 \times T$ size, 105, from which two outputs $X_{N,0}$ and $X_{N,1}$ of $1 \times T$ size are obtained, these outputs being the voice channel and noise channel, respectively.

[0020] These steps are iterated for each new frame. The parameters are trained using a database of sounds. The drawback of this method is that it does not make use of the information contained in the preceding frames when processing the current frame. This especially leads to a low processing quality and high processing latency, because of the length of the frames..

[0021] One of the objectives of the present invention is to provide a method and device allowing voices to be separated from background noise in an audio stream, or voice in an audio stream to be denoised, especially by taking into account information obtained from preceding frames. This allows performance and processing latency to be improved. The method thus allows propagation of “global information” on the signal, its update and its exploitation from frame to frame.

[0022] The invention relates to a method for separating in real time voice from noise in an audio signal received via a receiver equipped with an audio sensor, characterized in that it comprises at least the following steps:

- the received audio stream is separated into N frames X_N ,
- a tensor containing information on the entirety of the audio stream is associated with each frame X_N ,
- the frame X_N is transmitted to a first module M_1 , which generates a signal $X_N^{(1)}$,
- the tensor I_{N-1} obtained in the preceding step of processing the frame $X_N - 1$ is transmitted to a module M_3 ,

- the module M_3 receives as input a signal $X_N^{(2)}$, that is the result of the conversion of the signal $X_N^{(1)}$ by a module M_2 and concatenates $X_N^{(2)}$ and I_N in order to generate a signal $X_N^{(3)}$ of $2F \times T$ size,
- the signal $X_N^{(3)}$ is transmitted to a module M_4 in order to generate a signal $X_N^{(4)}$ that is combined with the signal $X_N^{(1)}$,
- the signal resulting from the combination is decoded by a decoder M_5 in order to generate a voice first signal $X_{N,0}$ and a second signal $X_{N,1}$.

[0023] To process a frame N , the frame $N - 1$ is assumed to have been processed beforehand and the quantities resulting from this processing to have been stored. For frame 0, I_0 is identically zero.

[0024] The invention also relates to a device for separating voice from noise in an audio signal received via a receiver equipped with an audio sensor, characterized in that it comprises at least the following elements:

- a first module M_1 that receives frames of a signal containing voice and noise,
- the first module having an output connected to a second module M_2 configured to generate a signal transmitted to a third module M_3 that receives a tensor value associated with a preceding frame $X_N - 1$ with a view to generating a tensor I_N associated with the current frame and a signal $X_N^{(3)}$ of $2F \times T$ size.

The module M_3 inserted between the module M_2 and the module M_4 receives as input a tensor of same size as that delivered as output from the module M_2 and delivers as output a tensor of same size as that received as input by the module M_4 . An additional input $I_N - 1$ is received as input by the module M_3 for the purposes of processing frame number N , and the module M_3 delivers, by way of additional output, the tensor I_N .

- a module M_4 that combines the signal $X_N^{(3)}$ and the signal $X_N^{(1)}$ in order to generate a signal $X_N^{(4)}$,
- a decoder M_5 configured to generate a voice first signal $X_{N,0}$ and a second signal $X_{N,1}$ from the signal $X_N^{(4)}$.

[0025] Other features, details and advantages of the invention will become apparent on reading the description, which is with reference to the appended drawings, which are given by way of non-limiting example and which show, respectively:

[0026] [Fig.1], an illustration of the prior art,

[0027] [Fig.2], an example of a system allowing the method according to invention to be implemented,

[0028] [Fig.3], an illustration of the steps implemented by the method according to the invention.

[0029] Figure 2 illustrates an example of a device allowing the method according to the invention to be implemented.

[0030] The signal the one or more voices of which must be extracted (separated) from the noise contained in the audio stream is received via an audio sensor 10. The audio sensor is connected to a set of pieces of equipment or hardware modules 20 that are configured to separate voice from noise, which modules will be described in detail with reference to Figure 3.

[0031] Figure 3 illustrates a first variant embodiment for separating a voice from noise in an audio signal, the processing operations being performed by the set 20. This separation is performed in real-time. Modules that are similar to those in the schematic of Figure 1 have been given the same references. The set in addition comprises a module M_3 the function of which is described in detail below.

[0032] The audio signal received via the sensor is, in a first step, separated into N frames $X_1, \dots, X_N$. With each frame X_N is associated a tensor I_N that is of a constant size that is independent of the index of the frame. The method will update the value of the tensor I_N from frame to frame and use X_N and I_N conjointly to estimate $X_{N,0}$ and $X_{N,1}$.

[0033] The frame X_N is transmitted to a first module M_1 , 100, which generates a signal $X_N^{(1)}$. The tensor I_{N-1} obtained in the preceding step of processing the frame X_{N-1} is transmitted to a module M_3 , 201.

[0034] M_3 generates a tensor I_N , 202, which will be used in the processing of the frame X_{N+1} .

[0035] The coder M_3 receives as input a signal $X_N^{(2)}$, 203, that is the result of the conversion of the signal $X_N^{(1)}$ by a module M_2 and concatenates $X_N^{(2)}$ and I_N , in order to generate a signal $X_N^{(3)}$ of $2F \times T$ size,

[0036] $M_3: (X_N^{(2)}, I_{N-1}) \rightarrow (X_N^{(3)}, I_N)$.

[0037] The signal $X_N^{(3)}$, 204, is transmitted to a module M_4 in order to generate a signal $X_N^{(4)}$ that is combined, 104, with the signal $X_N^{(1)}$, and the signal resulting from this combination is decoded by a decoder M_5 , 105, in order to generate a voice first signal $X_{N,0}$ and a second signal $X_{N,1}$.

[0038] According to one embodiment, the steps implemented by the method according to invention are the following:

[0039] For any N , I_N is of $F \times F$ size.

[0040] A_N is an $F \times F$ tensor defined by $A_N = \left(\frac{X_N^{(2)} \cdot (X_N^{(2)})^t}{\sqrt{T}} \right)$.

a. $X_N^{(2)} \cdot (X_N^{(2)})^t$ is the matrix product of $X_N^{(2)}$ and its transpose.

[0041] $I_N = I_{N-1} + \lambda(A_N - I_{N-1})$ with λ a gain factor 0 and 1 given by the user.

[0042] $B_N = \text{Softmax}(I_{N-1})$

a. The softmax function is conventionally used in machine learning; with a vector of K numbers, $(v_1 \dots v_K)$, it associates a vector of K numbers, $(w_1 \dots w_K)$, with, for every number $w_k = \frac{\exp(v_k)}{\sum_{l=1}^K \exp(v_l)}$.

b. To compute B_N , the softmax function is applied independently to all the lines of I_N .

[0043] $C_N = B_N \cdot X_N^{(2)}$ is the matrix product of B_N and $X_N^{(2)}$; its size is $F \times T$,

[0044] $X_N^{(3)}$, i.e. the concatenation of $X_N^{(2)}$ and C_N , is of $2F \times T$ size.

[0045] The method and device according to the invention allow voice to be separated from noise in an audio signal received via a sensor in real-time without degrading the parameters specific to the voice.

[0046] The invention allows real-time operation with a controllable latency/quality compromise, allows the audio signal that contains no noise not be degraded, and allows noise to be enhanced in a signal containing no words (voice).

[0047] The method especially allows the audio signal of the speech to be preprocessed with a view to improving the quality of voice processing/enhancement (compression, analysis) blocks.

[0048] The addition to the processing chain of a module M_3 allows the quality with which a frame-by-frame strategy may be implemented to achieve processing in real-time to be improved.

CLAIMS

1. Method for separating in real-time voice from noise in an audio signal received via a receiver equipped with an audio sensor, characterized in that it comprises at least the following steps:

- the received audio stream is separated into N frames X_N ,
- a tensor containing information on the entirety of the audio stream is associated with each frame X_N ,
- the frame X_N is transmitted to a first module M_1 , (100), which generates a signal $X_N^{(1)}$,
- the tensor I_{N-1} obtained in the preceding step of processing the frame X_{N-1} is transmitted to a module M_3 , (201),
- the module M_3 receives as input a signal $X_N^{(2)}$, (203), that is the result of the conversion of the signal $X_N^{(1)}$ by a module M_2 and concatenates $X_N^{(2)}$ and I_N , in order to generate a signal $X_N^{(3)}$ of $2F \times T$ size,
- the signal $X_N^{(3)}$, (204), is transmitted to a module M_4 in order to generate a signal $X_N^{(4)}$ that is combined, (104), with the signal $X_N^{(1)}$,
- the signal resulting from the combination is decoded by a decoder M_5 , (105), in order to generate a voice first signal $X_{N,0}$ and a second signal $X_{N,1}$.

2. Device for separating voice from noise in an audio signal received via a receiver equipped with an audio sensor, characterized in that it comprises at least the following elements:

- a first module M_1 that receives frames of a signal containing voice and noise,
- the first module having an output connected to a second module M_2 configured to generate a signal transmitted to a third module M_3 that receives a tensor value associated with a preceding frame X_{N-1} with a view to generating a tensor I_N associated with the current frame and a signal $X_N^{(3)}$ of $2F \times T$ size,
- a module M_4 that combines the signal $X_N^{(3)}$ and the signal $X_N^{(1)}$ in order to generate a signal $X_N^{(4)}$,
- a module (104) that combines the signal $X_N^{(4)}$ with the signal $X_N^{(1)}$ in order to generate a signal $X_N^{(5)}$,

- a decoder M_5 , (105) configured to generate a voice first signal $X_{N,0}$ and a second signal $X_{N,1}$ from the signal $X_N^{(5)}$.