観測信号の相関係数に基づく無音区間判別を利用した 騒音環境下での話者音声抽出

○樋口 佳奈^a 大隈 千春^a 石橋 孝昭^a

^a 熊本高等専門学校

1. はじめに

観測された混合信号のみを用いて原信号を推定するブラインド信号分離 (BSS: Blind Source Separation) において, 話者音声に含まれる無音区間を利用し, 観測信号の分布を回転させることで, 騒音内で発話された音声から雑音のみを除去する方法を提案する。提案法は, 短時間処理のために音声信号を短時間フレームに分割して, 目的話者音声を抽出する。このとき, フレーム内に無音区間が存在しないときには, 分布を回転できない問題が生じる。そこで, 分布の相関係数に基づいて, フレーム内の無音区間の有無を判別する。無音区間のないフレームには, 無音区間が存在するフレームのパラメータを使うことで目的音を抽出する。

2. ブラインド信号分離

ブラインド信号分離は, 原信号 $s_n(t)$ ($1, 2, \dots, N$) と伝達関数 a_{mn} が未知の条件下で, 観測された混合信号 $x_m(t)$ ($m = 1, 2, \dots, M$) のみを用いて原信号を推定する技術である。最も簡単な瞬時混合過程は

$$x_m(t) = \sum_{n=1}^N a_{mn} s_n(t) \quad (1)$$

のように表現される。混合信号 $x_m(t)$ を用いて, 原信号 $s_n(t)$ の推定値である分離信号 $y_n(t)$ は

$$y_n(t) = \sum_{m=1}^M w_{nm} x_m(t) \quad (2)$$

のように推定される。 w_{nm} は分離荷重である。

独立成分分析 (ICA: Independent Component Analysis) をはじめとして, これまで様々な分離法が提案されており, 原信号の数がマイクロホン数と等しい場合, 源信号はほぼ完全に復元できている。しかしながら, 多くの手法が逐次更新アルゴリズムであり処理に時間がかかるという問題がある。

3. 観測信号の分布の回転に基づく話者音声の抽出

著者らはこれまでに, 2 音源 2 マイクロホンの環境下で観測信号の同時分布を回転して話者音声を推定する方法を提案して

いる。さらに, その分離法を短時間フレーム処理できるように拡張した^{(1),(2)}。処理フレーム長が T のとき, l 番目の処理フレームの混合信号 $x_{m,l}(t)$ は

$$x_{m,l}(t) = [x_m((l-1)T), x_m((l-1)T+1), \dots, x_m(lT-1)] \quad (3)$$

のように表記できる。観測した混合信号の同時分布を Fig. 1 に示す。同時分布とは, 時間軸を無視して平面に全ての点をプロットした図のことである。Fig. 1 の横軸はチャンネル 1 で観測した混合信号の振幅, 縦軸はチャンネル 2 で観測した混合信号の振幅である。短時間フレーム内に無音区間があれば, 観測信号の同時分布は Fig. 1 の左上の処理フレームのように直線成分を持つ。このとき, 分布の相関係数は 1 に近づくことになる。

そこで, 処理フレーム l 番目の同時分布に対する相関係数を以下のように計算する。

$$r_l = \frac{\sum_{t=0}^{T-1} (x_{1,l}(t) - \overline{x_{1,l}})(x_{2,l}(t) - \overline{x_{2,l}})}{\sqrt{\sum_{t=0}^{T-1} (x_{1,l}(t) - \overline{x_{1,l}})^2} \sqrt{\sum_{t=0}^{T-1} (x_{2,l}(t) - \overline{x_{2,l}})^2}} \quad (4)$$

相関係数 r_l に基づいて相関の有無を判別し, 相関が 1 に近い場合には観測信号 $x_{m,l}(t)$ の分布に対する方位 $\phi_l(t)$ を

$$\phi_l(t) = \tan^{-1} \frac{x_{2,l}(t)}{x_{1,l}(t)} \quad (5)$$

Fig. 1: Joint distributions of observed signals.

Human speech extraction under noisy environment using silence interval determination based on correlation coefficients of observed signals
Kana Higuchi^a, Thiharu Ookuma^a, Takaaki Ishibashi^a
^aKumamoto Institute of Technology, Kumamoto college

Fig. 2: Histograms of observed signals.

のように計算する。

さらに、 $\phi_l(t)$ のヒストグラム $\text{hist}(\phi)$ を Fig. 2 のように求める。ヒストグラムは横軸に角度、縦軸に度数分布をとったグラフである。Fig. 2 から、相関が高いほど分布は一つの方位に集中していることがわかる。これは、話者音声には無音区間があるため、騒音しか存在しない頻度が高くなるからである。これに基づいて、ヒストグラム $\text{hist}(\phi)$ の最頻値 θ_l を式 (6) で求める。

$$\theta_l = \arg \max_{\phi} \text{hist}(\phi) \quad (6)$$

式 (6) より推定信号 $y_l(t)$ を

$$y_l(t) = \cos \theta_l x_{2,l}(t) - \sin \theta_l x_{1,l}(t) \quad (7)$$

のように推定し騒音環境下で観測された混合信号から l 番目の処理フレームの話者音声を抽出できる。

相関係数が 1 から離れた値のフレームの話者音声の推定には、すでに処理を行い相関係数 r_l が 1 に近いと判断した直前のフレームのパラメータを使うことで話者音声を抽出する。推定信号は話者音声を推定しているので、各処理フレームの推定信号をそのまま並べるだけで、原信号が復元される。また、相関係数 r_l を短時間フレームごとに計算しているため話者が移動するなど観測途中で環境が変わった場合の目的音推定も期待できる。

4. シミュレーション

音声データベースの 6 秒間の話者音声と自動車エンジン音を用いてシミュレーションを実施した。音源のサンプリング周波数は 8[kHz]、分解能は 16[bit] とした。20 パターンの混合信号を用いてシミュレーションした結果、全ての処理フレームで騒音が除去された話者音声を抽出できた。Fig. 3 は観測信号を 6 分割して 1 秒間の処理フレームのときのシミュレーション結果

Fig. 3: Simulation result.

である。(a) の $s_1(t)$ は話者音声、 $s_2(t)$ は騒音である。この 2 つ音源の混合信号を使用し、原信号の推定を行なった。

(b) の信号の同時分布とヒストグラムは、それぞれ前述の Fig. 1 と Fig. 2 である。Fig. 1 および Fig. 2 より 1 番目の処理フレームや 6 番目の処理フレームの分布には直線成分がはっきりと現われ無音区間が存在している。これらのフレームのパラメータを利用して Fig. 3(c) の分離信号を推定した。(a) の音源 $s_1(t)$ と比較するとほとんど同じ話者音声のみを抽出できていることが確認できる。また、分割数を増やして 1 つのフレーム長が 0.01 秒のときにも話者音声抽出できることを確認した。

5. おわりに

騒音環境下で観測された混合信号から、観測信号の分布を回転して目的話者音声を抽出する方法を提案した。提案法は観測信号を短時間フレームに分割し、話者音声に含まれる無音区間を利用して分布を回転させる。このとき、無音区間が存在しないために分布を回転できない問題に対して、無音区間の有無を相関係数 r_l を用いて判別し、フレーム内に無音区間がない場合は無音区間のあるフレームのパラメータを使えるようにした。この処理により、全てのフレームから目的音の抽出が行なえる。本手法は、フレーム長が 0.01 秒のときにも処理が可能のため時間遅れが発生せず、発話の途中に音源が移動して環境が変わったときの目的音推定も期待できる。

文献

- (1) 樋口佳奈, 石橋孝昭, 大隈千春, 葉山清輝, 五反田博, “無音区間を利用した騒音環境下での話者音声の抽出,” 第 23 回電子情報通信学会九州支部学生会講演会論文集, D-29, 2015.
- (2) 樋口佳奈, 石橋孝昭, 大隈千春, “無音区間を利用した短時間処理による騒音環境下での目的音抽出,” 2016 年日本生体医工学会九州支部学術講演会論文集, p.42, 2015.