4U-2

マイク数以上の同時発話分離のための調波・非調波音源モデルの検討

平澤 恭治　安良岡 直希　高橋 徹　尾形 哲也　奥乃 博

京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

近年スマートフォン上で動作する音声検索システムや音声エージェントなどが公開され, 一般ユーザにも音声入力システムが身近なものとなってきた. 現在の多くのシステムは接話マイクを通じて入力を得るが, 将来我々の周囲にある様々な機器が音声入力で動作する際には, 身体的拘束のない非接話マイクを通じてユーザー発話を得ることが望ましい. ただし, 非接話マイクの観測音は背景雑音や他話者の音声など多数の非目的音との混合音となるので, それらの非目的音を抑圧または分離する必要がある. 特に, 人間の生活環境には無数の音源が存在するため, 多数の音源が存在する環境への対応が重要である.

そこで我々は, 音源数がマイク数を上回る “劣決定状況” における同時発話の分離に焦点を当て, 劣決定同時発話分離のための GMM 音源モデルを提案した [1]. このモデルは調波用の複数のガウシアンと, 非調波用の複数のガウシアンを持ち, それらの総和で音声スペクトルを表現する. しかし GMM 音源モデルの非調波モデルは調波のモデルに似せて設計されたに過ぎず, 非調波スペクトルを上手く表現できずに最終的な分離性能を低下させている可能性があった.

本稿では, 従来のモデルより高い表現能力を持つ 2 つの非調波振幅モデルを提案する. 具体的には, 従来のモデルからガウシアンが等間隔で並んでいるという制約を取り除き, 音源・時間フレームごとにガウシアンの中心周波数を独立にしたモデル (モデル 1) と, 非調波振幅スペクトルをより正確に表現するため, ガウシアンではなくケプストラムから計算されるスペクトル包絡によって非調波成分を表現するモデル (モデル 2) を提案する.

2. 音源モデルを用いた音源分離

2.1 問題設定

本稿で扱う問題は論文 [1] と同様, 以下のように書ける.

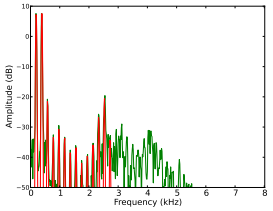
入力　J 個の同時発話を混合した I チャンネル音: $x_{i,fn}$
出力　J 個の推定された音源信号: $\hat{s}_{j,fn}$
仮定　劣決定状況, 線形時不変な畳み込み混合

2.2 GMM 音源モデル

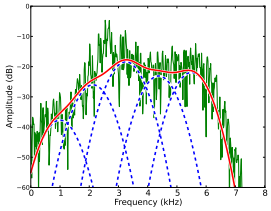
まず初めに論文 [1] で提案した GMM 音源モデルを説明する. 調波音を単音とその整数倍音の和だと考えると, 調波音をガウス窓を用いてフーリエ変換することにより, 振幅スペクトル上には鋭いガウシアンが等間隔で現れる [2]. このことから, 時間フレーム n での話者 j の基本周波数 $F^H_{0,j,n}$, 周波数ビン番号 f, 調波倍音番号 m_h を用いると, 調波モデルは以下のように定式化できる (図 1(a) 赤線).

$$\hat{s}^H_{j,fn} = \sum_{m_h} p^H_{j,n,m_h} \exp\left(-\frac{(f' - m_h F^H_{0,j,n})^2}{2\sigma^{H2}}\right) \quad (1)$$

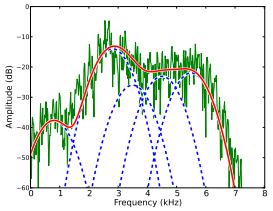
A Study About Harmonic and Nonharmonic Sound Source Model for Separating Simultaneous Utterances with Less Number of Microphones: *Yasuharu Hirasawa, Naoki Yasuraoka, Toru Takahashi, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto Univ.)*



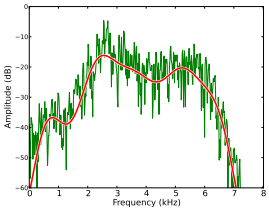
(a) 調波モデル



(b) 非調波モデル 0: 等間隔



(c) 非調波モデル 1: 非等間隔



(d) 非調波モデル 2: ケプストラム

図 1: 調波モデルと各非調波モデルのイメージ図

ここで σ^{H2} はスペクトル上のガウシアンの分散で, これはガウス窓の分散から計算される. f' は f 番目の周波数ビンの中心周波数であり, p^H_{j,n,m_h} は各ガウシアンの複素振幅で, 振幅成分と位相成分を共に含んでいる.

一方, 非調波用の信号は周波数ビンごとに位相が異なるため, まず $\hat{s}^N_{j,fn} = |\hat{s}^N_{j,fn}|\phi^N_{j,fn}$ として振幅成分 $|\hat{s}^N_{j,fn}|$ と位相成分 $\phi^N_{j,fn}$ に分離する. 非調波成分は調波成分に比べて滑らかな振幅スペクトルを持つので, 大きな分散を持った非調波用ガウシアンを周波数軸上で等間隔に並べて $|\hat{s}^N_{j,fn}|$ を表現する. 非調波のための (擬似的な) 基本周波数 F^N_0 と非調波倍音番号 m_n を用いると, 非調波振幅モデルは以下のように定式化できる.

$$|\hat{s}^N_{j,fn}| = \sum_{m_n} p^N_{j,n,m_n} \exp\left(-\frac{(f' - m_n F^N_0)^2}{2\sigma^{N2}}\right) \quad (2)$$

ここで σ^{N2} は非調波用ガウシアンの分散で, p^N_{j,n,m_n} は各ガウシアンの実数振幅である. 図 1(b) はガウシアンの数を減らしたイメージ図であり, 青点線が各ガウシアン, 赤実線がそれらの和で $|\hat{s}^N_{j,fn}|$ を表している. 本稿では以下この非調波振幅モデルをモデル 0 と呼ぶ.

なお最終的な音源モデルは $\hat{s}_{j,fn} = \hat{s}^H_{j,fn} + \hat{s}^N_{j,fn}$ となる.

2.3 音源モデルを用いた音源分離

音源モデル $\hat{s}_{j,fn}$ と混合行列 $\{a_{ij,f}\}$ を用いると, マイク i での観測音の推定値 $\hat{x}_{i,fn}$ を計算できる. 実際に観測された混合音 $x_{i,fn}$ とその推定値 $\hat{x}_{i,fn}$ の二乗誤差の最小化を考えると, コスト関数は以下のように表せる.

$$C = \sum_{ifn} |x_{i,fn} - \hat{x}_{i,fn}|^2 = \sum_{ifn} |x_{i,fn} - \sum_j a_{ij,f}\hat{s}_{j,fn}|^2 \quad (3)$$

論文 [1] では, このコスト関数に対して補助関数法 [3] を用い, 解析的に更新則を導出した. 本稿ではスペースの都

合から導出の詳細は省略する.

以下では非調波振幅モデル $|\hat{s}^N_{j,fn}|$ を変更し, より高精度な音源分離の実現を目指す. その場合でも調波モデル $\hat{s}^H_{j,fn}$, 非調波の位相 $\phi^N_{j,fn}$, コスト関数 C など, $|\hat{s}^N_{j,fn}|$ の数式表現以外は上で述べたものを使用する.

3. 提案する非調波振幅モデル

3.1 非等間隔ガウシアンによるモデル (モデル 1)

非調波振幅モデル 0 では, 調波用モデルとの差分を少なくするため, 非調波用ガウシアンを等間隔に並べていた. しかし実際の音声スペクトルでは非調波成分のピーク位置は任意である. そこで非調波ガウシアンが独立に移動できるモデルを考えると, 非調波振幅モデルは以下の式により定式化できる.

$$|\hat{s}^N_{j,fn}| = \sum_{m_n} p^N_{j,n,m_n} \exp\left(-\frac{(f' - F^N_{0,j,n,m_n})^2}{2\sigma^{N2}}\right) \quad (4)$$

モデル 0 ではガウシアンの位置が $m_n F^N_0$ となっていたが, このモデルでは F^N_{0,j,n,m_n} と変更され, 音源や時間フレームに応じて移動できるようになった (図 1(c)).

このモデルを用いた際の F^N_{0,j,n,m_n} の更新則は, 論文 [1] 中での $F^H_{0,j,n}$ の更新則と同様に求まり, 以下のようになる.

$$F^N_{0,j,n,m_n} = \sum_{if} \nu_{ij,fn,m_n} e^{-\gamma^N_{j,fn,m_n} f} / \sum_{if} \nu_{ij,fn,m_n} e^{-\gamma^N_{j,fn,m_n}} \quad (5)$$

$$\nu_{ij,fn,m_n} = {\beta^N_{ij,fn,m_n}}^{-1} \Re e\left[\alpha^N_{ij,fn,m_n} x_{i,fn} a^*_{ij,f} p^N_{j,n,m_n} \phi^{N*}_{j,fn}\right] \quad (6)$$

ここで上付き文字の * は複素共役, $\Re e[..]$ は複素数の実部を表す. なお α^N_{ij,fn,m_n} と γ^N_{j,fn,m_n} は補助関数法を用いた際に導入された補助変数で, β^N_{ij,fn,m_n} は補助関数のパラメータである. p^N_{j,n,m_n} など F^N_{0,j,n,m_n} 以外のパラメータの更新則は論文 [1] と同様に導出する.

3.2 ケプストラムを用いたモデル (モデル 2)

モデル 0 やモデル 1 では, ガウシアンの和を用いて非調波スペクトルを表現していた. モデル 2 ではケプストラムから計算されるスペクトル包絡を用い, より表現力の高い非調波振幅モデルを実現する (図 1(d) 赤線). この時非調波振幅モデルはケプストラム係数 $c_{j,n,k}$ を用いて次のように定式化できる.

$$|\hat{s}^N_{j,fn}| = h^N_{j,fn} = \exp\left\{c_{j,n,0} + 2\sum_k c_{j,n,k} \cos(\omega k)\right\} \quad (7)$$

ここでフレーム長 W を用いて $\omega = 2\pi f / W$ である.

モデル 0 やモデル 1 では全てのパラメータ更新則が解析的に求められたが, このモデルでは解析的な導出が困難である. そこで本稿では各音源 j, 各時間フレーム n ごとに以下の手順で $c_{j,n,k}$ の更新を行った.

1. 次の式を用いて, 目標振幅スペクトルを計算する

$$\bar{h}^N_{j,fn} = \left| \frac{\sum_i {\beta^N_{ij,fn,m_n}}^{-1} \alpha^N_{ij,fn,m_n} x_{i,fn} a^*_{ij,f} \phi^{N*}_{j,fn}}{\sum_i {\beta^N_{ij,fn,m_n}}^{-1} |a_{ij,f}|^2} \right| \quad (8)$$

2. 目標振幅スペクトルの対数を取り, 逆フーリエ変換した後, 低次の係数から順に K 要素を $c_{j,n,k}$ とする.

モデル 1 と同じく, その他のパラメータの更新則は論文 [1] と同様に導出する.

4. 実験

本稿で提案した非調波音源モデルの有効性を確認するため, SiSEC 2011 で配布されている複数話者混合音声の分離実験を行った. 実験に用いたデータは女性 3 話者の同時発話を残響 130ms の環境でステレオ録音したものである. 分離結果の評価には, 全体の分離性能を示す SDR, 時間的・空間的歪みを示す ISR, 他話者からの漏れノイズの少なさを示す SIR, その他の歪みを示す SAR の 4 尺度 [4] を用いた. 実験に用いた定数パラメータの設定値は表 1 の通りである. なお, 非調波成分の分離性能を正しく比較するため, 基本周波数 $F^H_{0,j,n}$ の初期値には手動で作成したオラクルデータを与えた.

表 1: 実験時に用いた定数パラメータ

サンプリング周波数	16kHz
STFT 窓長: W	1024 点 (64ms)
STFT シフト幅	256 点 (16ms)
$I/J/F/N$	2 / 3 / 513 / 622
$M_H/M_N/K$	40 / 34 / 10

表 2: 分離実験の結果 (dB)

非調波振幅モデル	SDR	ISR	SIR	SAR
0: 論文 [1] のもの	6.3	11.6	11.9	6.9
1: 非等間隔ガウシアン	6.5	12.6	13.2	6.8
2: ケプストラム	6.7	12.1	12.6	7.3

表 2 に, 最終分離結果に対する 4 尺度の評価値を示す. なおこの分離結果は, 目的話者の $\hat{s}_{j,fn}$ を信号, 他話者の $\hat{s}_{j,fn}$ を雑音とみなし, 観測 $x_{i,fn}$ に対して Wiener フィルタを適用したものである. 全体の分離性能を評価する SDR を見ると, 論文 [1] で提案したモデル 0 に比べ, モデル 1 で 0.2dB, モデル 2 で 0.4dB の改善が見られた. また, モデル 2 では全ての尺度で性能の向上が確認された.

$|\hat{s}^N_{j,fn}|$ を表すためのパラメータ数は, モデル 0 で 35 $(1 + M_N)$, モデル 1 で 68 $(M_N + M_N)$ だったのに対し, モデル 2 で 10 (K) と大きく減少している. 自由度は大きく低下したものの性能が向上したのは, ケプストラムから求まるスペクトル包絡の表現力の高さだと考える. なお, これ以上ケプストラム係数の個数 K を増やしても性能は向上しなかった. これは, 表現力が高くなりすぎると分離中に生じる歪みにも適応し始めるためと考えられる.

5. おわりに

本稿では劣決定状況における同時発話分離のために, 新たに 2 種類の非調波振幅モデルを提案した. 提案したモデルのうちケプストラムを用いたものは, SiSEC で使用された劣決定同時発話音声の分離実験において, 従来の GMM 音源モデルに対し SDR で 0.4dB の向上を実現した. 今後はより良いパラメータ初期化手法やパラメータ更新順序, 他の音源分離手法との統合等を検討したい.

本研究の一部は科研費 (S), GCOE, 日仏研究交流の支援を受けた.

参考文献

[1] Y. Hirasawa, N. Yasuraoka, T. Takahashi, T. Ogata, and H.G. Okuno: "A GMM Sound Source Model for Blind Speech Separation in Underdetermined Conditions", In *Proc. of LVA/ICA 2012*, to appear.

[2] H. Kameoka, N. Ono, and S. Sagayama: "Auxiliary Function Approach to Parameter Estimation of Constrained Sinusoidal Model for Monaural Speech Separation", In *Proc. of ICASSP 2008*, pp. 29–32.

[3] D.D. Lee and H.S. Seung: "Algorithms for non-negative matrix factorization", *Advances in neural information processing systems*, Vol. 13, pp. 556–562, 2001.

[4] E. Vincent, H. Sawada, P. Bofill, S. Makino and J.P. Rosca: "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results", *Independent Component Analysis and Signal Separation*, pp 552–559, 2007.