話者混在音声からの fo 軌跡の分離 *

○森　大毅 (宇都宮大)

1　はじめに

既存の会話コーパスの 100 倍規模となる「日本語日常会話コーパス (CEJC)」の登場によって、会話音声の韻律の研究は新たな段階に入ろうとしている。しかし、各話者の音声は音響的に分離されておらず、互いの音声の回り込みが無視できない。このため、複数話者の同時発話区間においては fo (基本周波数) が信頼できず、定量的な韻律研究の障害となっている。

本研究の目的は、日常的な生活環境で収録した話者混在音声から、会話音声の韻律のモデル化に必要な fo 情報を話者ごとに分離する技術の確立である。これは一般化した fo 推定問題と考えられるが、既存の fo 推定法が各時刻でただ 1 つの fo が存在することを前提としているのに対し、本研究では目標話者ごとに異なる fo が存在するというモデルになっているのが特徴である。

2　問題設定

CEJC は非同期の多チャンネル信号であり、アドホックマイクロホンアレー技術等により多チャンネルの情報を活用した音声強調が可能と考えられる。本研究では、問題の簡単化のため 1 チャンネル信号を仮定する。

fo の分離対象となる話者混在音声では、話者の同一性 (ID) は既知であると仮定する。さらに、各話者の発話区間も既知であるとする。これは、CEJC のような発話区間ラベルを持つコーパスにおいては自然な仮定である。

推定した fo 軌跡の精度を評価するためには正解となる fo 情報が必要であるが、CEJC はそのような情報を提供していない。本稿では、模擬話者混在音声を用いる。具体的には、2 話者のクリーン音声を人工的に重畳させることで話者混在音声を模擬し、元音声から推定した fo 軌跡を正解として評価する。

3　モデル構造

提案モデルの構造を図 1 に示す。モデルへの入力は話者混在音声のスペクトログラムであり、

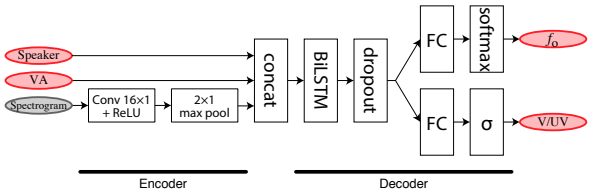

Fig. 1　提案モデルの構造

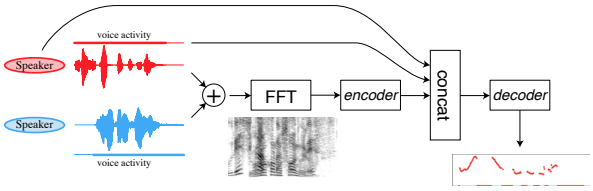

Fig. 2　1 人目の話者に対する損失計算

16 × 1 畳み込み層 (32 チャンネル)、2 × 1 max プーリング層を経て特徴抽出される。このベクトルは、目的話者の発話区間情報 (1/0) および目的話者の埋め込みベクトルと結合された後、fo 軌跡の時間変化パターンを作り出す双方向 LSTM(隠れユニット数 512) に送られ、量子化された fo および有声／無声情報 (1/0) を出力する。過去に提案されている音声分離のためのニューラルネット [1] と異なり、このモデルは 1 度に 1 話者の fo 軌跡だけを出力する。各話者の fo 軌跡は、それぞれの話者に対応する話者埋め込みを入力に添えてネットワークを駆動することで 1 つずつ得られる。

4　ネットワークの学習

話者混在音声を模擬するため、異なる 2 話者の 2 発話を訓練データからランダムに抽出し、前後にシフトして重畳する。このランダムペアリングとシフトは訓練エポック毎に行う。

損失は重畳音声の話者全員に対して積算する。重畳音声のスペクトログラムに 1 人目の話者の埋め込みベクトルと発話区間情報を結合して fo 軌跡を推定し、その話者の正解 fo データを教師信号として損失を計算する (図 2)。これをほかの話者に対しても繰り返し、最後に全話者の損失を積算する。

---

* fo contour separation from mixed speech.
by MORI, Hiroki (Utsunomiya Univ.)

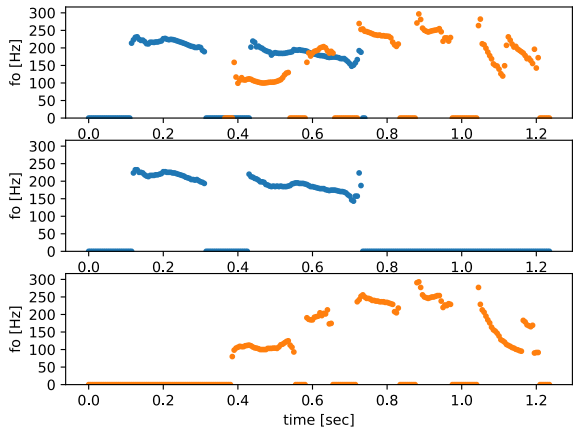

Fig. 3　正解／推定 fo 軌跡 (1)

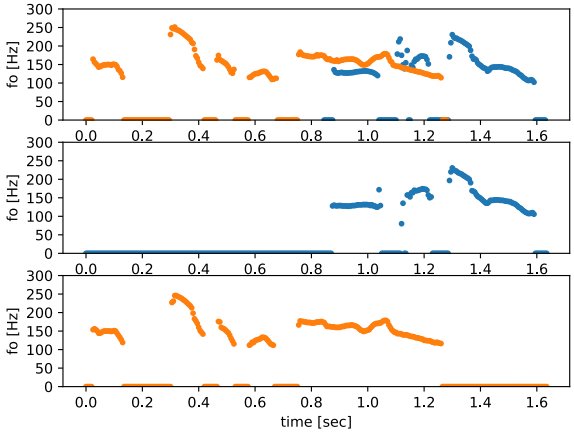

Fig. 4　正解／推定 fo 軌跡 (2)

5　実験

クリーン音声のコーパスとして、日本語話し言葉コーパス (CSJ) の模擬講演音声を用いた。フィラーや言い淀みなどの非流暢性を含む発話は除いた。次に、これを訓練用セット (344 名, 98967 IPUs)、適応用セット (訓練用セットに含まれない 20 名, 7555 IPUs)、テストセット (適応用セットと同じ 20 名, 3238 IPUs) に分割した。

正解 fo データは YangSaf [2] により 5 ms 間隔で求め、fo の値は 80 Hz から 600 Hz までの 255 段階に量子化した。

今回は、テストセットの話者は埋め込みが既知であるとの前提で実験を行った。テストセットの話者埋め込みは、訓練用セットと適応用セットの和集合から話者埋め込みまで含めてモデルを学習し、そこから取り出すことで得た。以降の実験では、適応用セットはモデル学習そのものには使用しない。

図 3, 4 に分離した fo 軌跡の例を示す。上段が正解、中・下段が各話者に対して推定された fo 軌跡である。どちらの例に対しても fo 軌跡の推定が精度良く行われていることがわかる。図 3 の例では女性話者と男性話者の fo 軌跡が交差している箇所があるが、分離は総じてうまく行っている。1.1 s から 1.2 s 付近では、男性話者の声質が creaky voice になっているため「正解」fo が 2 倍に誤っている部分があるが、提案法では fo 形状の高い表現能力のため、正しい fo に修正されている。

fo 軌跡分離能力を定量的に評価するため、有声無声誤り率 (VDE)、グロスエラー率 (GPE)、ファインエラー率 (FPE) を求めた。これらの誤り率は各話者の音声区間のみを対象として計算した。表 1 に結果を示す。「分離前」は、重畳音声に単純に YangSaf を適用した場合の誤り率である。音声が重畳していない区間では 2 条件の精度に大きな違いは無いが、重畳している区間では分離技術を用いない限り少なくとも片方の話者の fo 推定は確実に誤る。表に示されるように、提案法によって fo 軌跡を分離することで、有声無声誤りとグロスエラーを大きく減少させることができた。

Table 1　提案 fo 分離法の定量評価

	VDE (%)	GPE (%)	FPE (st)
分離前	10.72	17.50	0.33
提案法	3.99	6.56	0.34

6　おわりに

話者混在音声からの fo 軌跡の分離法を提案した。提案法では、CNN と RNN から成るニューラルネットを、人工的に重畳させた音声のスペクトログラムから指定した話者の fo 軌跡を推定するように訓練する。原理的には、この方法は話者が増えても対応可能である。

このモデルを CEJC 等のコーパスに実際に適用するためには、対象話者の埋め込みを知る必要がある。未知の話者に対して信頼できる話者埋め込みを推定することが今後の課題の 1 つである。また、CEJC のような実環境下での音声に対する fo 推定の有効性を検証する必要がある。

謝辞 本研究は JSPS 科研費 19H01252 の助成を受けている。

参考文献

[1] Kolbæk et al., IEEE/ACM TASLP, **25**, 1901–1913, 2017.

[2] Kawahara et al., Proc. SSW9, 238–245, 2016.