話者ダイアライゼーションにおける相づち認識の精度向上

Improved accuracy of back-channeling recognition in speaker diarization

後藤　大喜
Daiki Goto
広島市立大学　言語音声メディア工学研究室
Language and Speech Research Laboratory, Hiroshima City University

概要　本研究では，日本語音声を話者ダイアライゼーションした際の精度の向上を目指す．日本語話者特有の表現である短い相づちは話者ダイアライゼーションにおいて認識率が低く精度を落とす原因となっていると考えられる．本研究ではこれを改良するために手法の検討と実験を行った．また，実際に起こりうるエラーの種類についても考察を行い，今後どのような手法を用いてエラーを解消していくかについても検討する．

1　はじめに

昨今，コロナウイルスの影響に伴い会議等がオンラインで行われることが多く，これまでのような議事録作成が難しくなった．話者ダイアライゼーション（Speaker Diarization）とは，会議やラジオなどの複数の話者が話している音声において，「いつ」，「だれが」話しているかを推定する技術である．実際に実行すると，一人の話者が長く話し続ける部分は精度高く話者分類できる．しかし，聞き手が細かく相づちを打っている部分は話者の交替が多いためか精度が落ちる．日本語話者の相づちは，英語話者のそれと比べると約 6 倍多いというデータがある．さらに本実験で用いた音声データの相づちの長さは平均約 0.5 秒と非常に短い．そこで本研究では相づち認識の精度を向上させることが話者ダイアライゼーションの精度の向上につながると考え，相づち認識の精度を上げるために実験を行った．

2　提案手法

相づちの認識精度を向上させるために窓枠サイズとオーバーラップ率を変化させた．

窓枠サイズとは，話者ダイアライゼーション実行時に音声を切り出す際の音声の長さのことである．窓枠サイズを小さくすることによって相づちのような通常よりも短い発話の認識精度の向上を目指した．

オーバーラップ率は切り出した際に隣り合う音声同士の重なる部分の割合のことである．値を大きくすることでデータを周期的に細かく見ることができるため性能の向上が見込まれる．

3　実験内容

本研究では，相づち認識の精度向上を上げることで話者ダイアライゼーションの精度も向上すると考え実験を行った．話者を認識・分類する際の音声特徴量には x-vector，クラスタリング手法には，スペクトラルクラスタリングを用いた．本研究で用いたシステムの概要を図 1 に示す．

音声データには，自然な会話音声を使用したいため，友人同士の雑談を収録している「千葉大学 3 人会話コーパス」のデータを使用する．

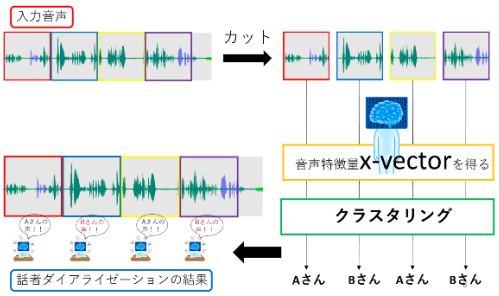

図 1:システムの概要

3.1　x-vector

x-vector とは，Deep Neural Network(DNN)から計算される音声特徴ベクトルであり，同様に音声特徴ベクトルである i-vector や d-vector と比べ新しく性能が良いとされている[1]．今回使用したモデルの学習には大量の英語の音声データが 3 秒程度にカットされ用いられている．

3.2　スペクトラルクラスタリング

スペクトラルクラスタリングとは，教師なし学習に分類されるクラスタリング手法のひとつである．k-means などのようなクラスタの中心からの距離でなくデータ同士の連結性に注目できるため他の方法ではうまくクラスタリングできなかったデータをきれいにクラスタリングできることがある．先行研究で話者ダイアライゼーションにおいては最も適しているクラスタリング手法とされている[2]．

3.3　実験

本実験では，切り出す音声の長さである窓枠サイズを 1.5 秒，1.0 秒，0.75 秒，0.5 秒に，隣り合う窓枠の重なり度合いを示すオーバーラップ率を 25%，50%，75%にそれぞれ変更して実験を行った．

3.4　追加実験

また，エラーの一種である誤検出を抑制するために音声の無音区間を信号のパワーとゼロ交差数の閾値から求める音声区間検出(VAD)によって求めた．VAD 無しでの結果とありでの結果を比較する実験を追加で行った．なお，窓枠サイズは 1.5 秒，オーバーラップ率は 50%とした．

3.5　評価手法

実験の評価には，Diarization Error Rate(DER）と相づちに注目したエラー率を用いた．　DER は一般的な話者ダイアライゼーションの評価指標である．求める式を(1)に，エラーの種類を説明する図を図 2 に示す．

$$\mathrm{DER} = \frac{False\ Alarm + \ Miss\ + Confusion}{Total} \cdots (1)$$

ここでの False Alarm は，誤検出した部分を表し，Miss は，相づち部分のラベルを非音声と分類した部分，Confusion は分類結果が別の人と間違えた部分の秒数を表す．Total は相づち部分のすべてのラベル合計時間を表す．相づちに注目したエラー率は相づち認識の精度を測定するための指標である．求める式を(2)に，エラーの種類を説明する図を図 3 に示す．

$$\text{相づちに注目したエラー率} = \frac{Miss\ + Confusion}{Total} \cdots (2)$$

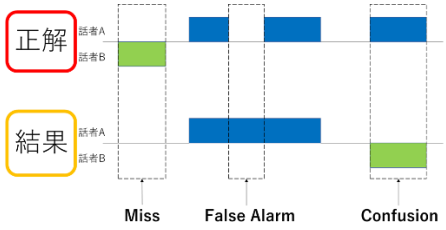

図 2:DER におけるエラーの種類

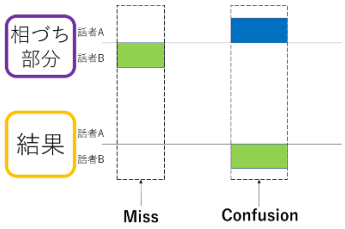

図 3:相づち部分のエラー率におけるエラーの種類

ここでの要素の示す意味は DER のときと同義である．どちらの指標もエラー率であるため，値が小さいほど精度は高い．

4　結果と考察

4.1　実験

「千葉大学 3 人会話コーパス」のデータのを，上で述べたパラメータで話者ダイアライゼーションした結果の DER と相づちに注目したエラー率を以下の表 1 と表 2 に示す．

表 1:DER の値[%]

窓枠サイズ[秒]	オーバーラップ率		
	25%	50%	75%
1.5	25.9	26.3	25.4
1.0	26	29.8	34.7
0.75	36.7	40.3	41
0.5	46.6	46.4	47.6

表 2:相づちに注目したエラー率の値[%]

窓枠サイズ[秒]	オーバーラップ率		
	25%	50%	75%
1.5	94.9	94	91.7
1.0	91.1	80.2	75.9
0.75	76.7	75.9	74.8
0.5	67.6	65.4	68.1

上の 2 つの表から窓枠サイズを小さくすると DER の値は悪化しているため，話者ダイアライゼーションの精度は悪化していることが分かる．ここではスペースの都合上掲載できていないが，エラー率の各要素の値を観察していくと，False Alarm や Miss の値はほとんど変化せず Confusion の値のみが変化していることがわかる．これは窓枠サイズが小さくなったことによって話者認識の難易度が上がったことで話者間違いが増えてしまったことが原因だと考える．本実験で用いた x-vector は 3 秒という窓枠サイズと比べ長い音声で学習されているモデルのためこのような結果になったと考える．

一方で相づちに注目したエラー率の値は良化している．これは，窓枠サイズを小さくしたことによって短い発話である相づちの認識率が上がったことを意味する．よって相づち認識の精度は向上したと言える．

また，オーバーラップ率を大きくすると，窓枠サイズが大きいときは，DER も相づちに注目したエラー率も良化した．これは，適度な長さで細かくデータを見ることができたからだと考える．窓枠サイズが小さいときは，データが細かすぎるためか DER の値は悪化した．一方で相づちに注目したエラー率の値は良くなっている．これは，窓枠サイズのときと同様，短い発話である相づちの認識率が上がったと考える．

4.2　追加実験

追加で行った VAD を行った際の結果と無しの時の結果 DER を表 3 に，相づちに注目したエラー率を表 4 に示す．

表 3:VAD の有無での DER の比較

	VADあり	VADなし
False Alarm[s]	36.149	42.253
Miss[s]	91.015	80.404
Confusion[s]	36.452	36.472
DER[%]	26.9	26.2

表 4:相づちに注目したエラー率の比較

	VADあり	VADなし
Miss[s]	8.004	7.895
Confusion[s]	57.568	57.677
相づちに注目したエラー率[%]	94.7	94.7

表 3 から DER の値は VAD 無しのほうがよかったことがわかる．検出ミス(= False Alarm)を減らすことには成功していが，Miss の値が増加してしまった．これは，VAD で無声と判定されえた部分の中にも正解ラベルがつけられていたり，そもそも今回用いた音声データの中に無声区間が少なかったため VAD の恩恵を受けにくかったのではと考える．実際に今回用いた音声データには 0.9 秒以上の無声区間は存在しなかった．そのため，VAD は不要と結論付けることはできないと考える．相づちに注目したエラー率にはほとんど変化が見られなかった．

5　まとめと今後の課題

本実験では，短い相づちの認識精度を高めることで話者ダイアライゼーション自体の精度も高まると考え実験を行った．しかし相づちの認識精度を高めると話者間違え(=Confusion)が多発してしまうため思ったように精度が上がらなかった．また，3 種類のエラーの中の 1 つである False Alarm を減らすための追加実験を行った．今後は他の 2 種類のエラーを解消するための実験を行いたいと考えている．Miss については，複数の話者が同時に発話している部分の処理が重要だと考えている．今のシステムでは 1 人の話者としか推定できないため複数の話者と判定するような学習器を作成したい. Confusion は, x-vector の学習法を検討したい. 学習データに英語音声だけでなく日本語音声を用い，短い発話にも対応できるように現在よりも短くカットして学習を行いたいと考えている．

6　参考文献

[1]Quan Wang, *et al* “Speaker diarization with LSTM” ICASSP 2018

[2] David Snyder, *et al*“X-vectors: robust DNN embeddings for speaker recgnition,” ICASSP 2018