会議の効率的な振返りを支援する話者クラスタリング技術

Speaker Clustering Technology to Assist in Performing Efficient Reviews of Speakers' Utterances Recorded at Meetings

木田　祐介　　丁　寧　　広畑　誠
■KIDA Yusuke　■DING Ning　■HIROHATA Makoto

スマートフォンやタブレットなどのモバイル端末の普及により，音声の録音がこれまでより手軽に行えるようになった。それに伴い，音声を録音するだけでなく，会議の効率的な振返りを支援する機能に対するニーズが高まっている。

東芝は，録音した音声に含まれる発話を話者ごとに分類する話者クラスタリング技術を開発した。この技術は，声の音色に関する特徴（音韻特徴）に加え，ステレオマイクを用いて推定した声の到来方向を必要に応じて併用することで，話者の分類精度を高めている。様々な会議で録音した音声を用いて実験を行い，従来に比べて高い精度で音声を話者ごとに分類できることを確認した。

The wide dissemination of mobile devices such as smartphones and tablets has greatly facilitated the recording and storage of voice data. Accordingly, demand has recently arisen for not only voice recording functions but also a function to assist in performing efficient reviews of the recorded utterances of speakers at meetings.

Toshiba has developed a speaker clustering technology that makes it possible to accurately classify utterances according to each speaker extracted from the recorded contents of a meeting. This is achieved by augmenting the extraction of timbre features with estimation of the speech arrival direction using stereo microphones as required. We have conducted evaluation experiments using speech recorded at various meetings and confirmed that this technology can classify utterances according to the corresponding speakers with higher accuracy than conventional methods.

1　まえがき

スマートフォンやタブレットなどのモバイル端末の普及により，音声の録音がこれまでより手軽に行えるようになった。それに伴い，音声を録音するだけでなく，会議の効率的な振返りを支援する機能に対するニーズが高まっている。

このような機能としては，例えば，話者を特定して発話を検索できる機能が挙げられる。また，近年の音声認識技術の発達により，話者と発話内容を記録する自動議事録生成機能にも高い注目が集まっている。これらの機能を実現するために重要な役割を担うのが，録音した音声に含まれる発話を話者ごとに分類する話者クラスタリング技術である。

話者クラスタリング技術は，これまでに様々な方法が提案されている。その多くは声の音色に関する特徴（音韻特徴）を用いる方法であり，特徴間の距離が近いものどうしを同じ話者による発話とみなして分類する。しかし，この方法では，声質の似ている話者を分類することが難しく，分類すべき話者の人数が多い場合に高い精度を得ることができなかった。

この問題を解決する方法として，別途推定した声の到来方向を用いる方法が提案されている[1]。この方法は，音韻特徴と方向の距離をそれぞれ計算し，両者に異なる重みを付けて足し合わせた距離を用いて話者を分類する。しかし，この方法では，同じ方向にいる話者どうしの発話が同じクラスタに分類されやすくなる課題があった。

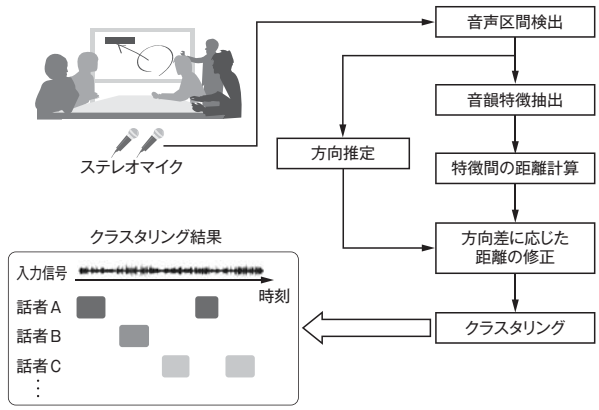

図1．話者クラスタリング技術の構成 ― 音韻特徴間で計算した距離を別途推定した方向の差によって修正し，修正した距離に基づいてクラスタリングする。
Configuration of speaker clustering technology

そこで東芝は，音韻特徴をベースに，ステレオマイクを用いて推定した声の到来方向を必要に応じて併用する新しい話者クラスタリング技術[2]を開発した。この技術は，同じ方向にいる話者など，ステレオマイクでは区別できない方向にいる話者どうしの発話に対しては混同を避けつつ，方向を生かして分類精度を高められる特長がある。様々な会議で録音した音声

を用いて実験を行い，従来の技術より高い精度で音声を話者ごとに分類できることを確認した。ここでは，この話者クラスタリング技術の概要と，その有効性を確認した実験結果について述べる。

2　話者クラスタリング技術

開発した話者クラスタリング技術の概要を**図1**に示す。音声の収録はステレオマイクで行う。はじめに，音声区間検出処理によって入力信号から人の発話していない区間を取り除き，残された区間で音韻特徴を抽出する。その後，異なる音韻特徴間の距離をそれぞれ計算し，別途推定した信号の到来方向の差に応じて距離を修正する。最後に，修正した距離に基づいてクラスタリングすることで，発話を話者ごとに分類した結果が得られる。以下，この処理の詳細を説明する。

2.1　音韻特徴の抽出

音韻特徴の抽出方法を**図2**に示す。音韻特徴を抽出するには，まず，入力信号の周波数特性を表した特徴量であるMFCC（Mel Frequency Cepstrum Coefficient）を求める。MFCCをそのまま音韻特徴として利用することもできるが，話者ごとに学習により統計的なモデルを求め，このモデルに対するMFCCのもっともらしさ（尤度（ゆうど））を用いることで，分類精度が高まることが知られている。そこで，話者交代点検出処理によって，連続してひとりの話者が発話した区間ごとに入力信号を区切り，その区間内に含まれるMFCCを用いた学習により統計モデルの一種であるGMM（Gaussian Mixture Model）を求める。その後，録音した音声からの学習で得られた全てのGMMに対してMFCCの尤度を算出し，これらを成分とするベクトル（尤度ベクトル）を音韻特徴として用いる。

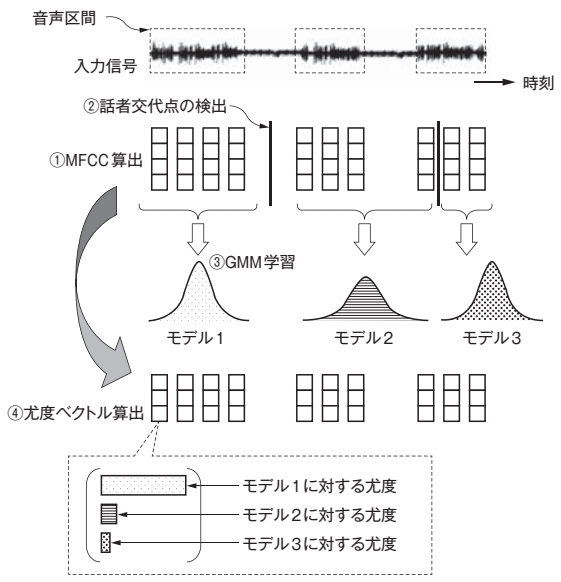

図2．音韻特徴の抽出方法 ― 話者交代点で区切った区間ごとにGMMを学習により求めた後，各GMMに対するMFCCの尤度を成分とする尤度ベクトルを算出することで音韻特徴を得る。

Method for extracting timbre features

2.2　ステレオマイクを用いた高速な方向推定

これまで，信号の到来方向を推定するためには，特殊な指向性を持つマイクを搭載したデバイスが必要であることや，処理が複雑で計算コストが大きいなどの問題があった。そこで，一般のステレオマイクを用いて高速に方向推定ができる方法を新たに開発した。

開発した方向推定のアルゴリズムを**図3**に示す。まず，二つのマイクに入力された信号の位相差を周波数ごとに計算し，図3(a)に示すような分布を得る。次に，以降の処理に必要な計算コストを抑えるために分布を量子化する（図3(b)）。量子化された分布は信号の到来方向に応じたパターンを描く。そこで，方向ごとに定まっている分布のパターンを理想分布（図3(c)）としてあらかじめ用意しておき，この中から量子化後の分布にもっとも近いものを選ぶことで方向を推定する。ここで，用意する理想分布のパターン数によって推定する方向の細かさを制御できるが，話者を識別するには3° の分解能で十分と考え，3° ごとに理想分布を用意した。

この方法は，分布どうしの比較だけで方向を推定するため，高速に処理できる。

2.3　音韻特徴間の距離計算と方向に応じた修正

二つの尤度ベクトルのユークリッド距離により音韻特徴間の距離を求めた後，推定した方向の差に応じて距離を修正する。二つの発話の方向差が小さい場合，それらは同じ話者から発せられた可能性があるが，同じ方向にいる別の話者から発せられた可能性もあるため，距離を大きくすることで分類精度を低下させるおそれがある。一方，二つの発話の方向差が大きい場合，それらは別の方向にいる話者から発せられた可能性

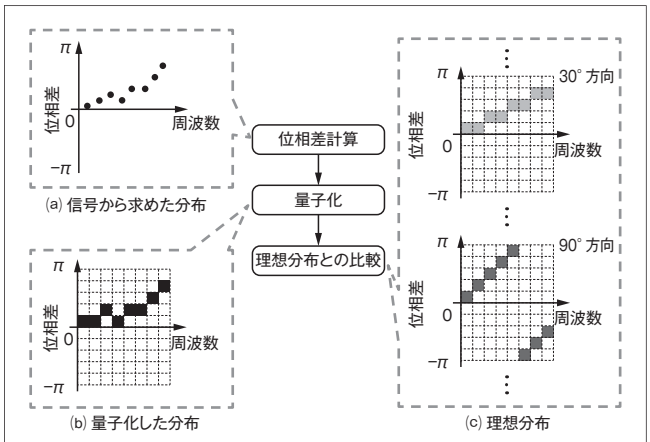

図3．方向推定のアルゴリズム ― 信号から求めた位相差の分布を量子化し，事前に用意した方向ごとの理想分布からもっとも近いものを選ぶことで方向を推定する。

Algorithm for direction of arrival (DOA) estimation

一般論文

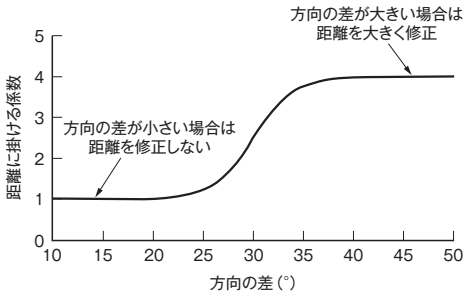

図4．シグモイド関数による距離の修正 — 方向の差が大きい場合に大きな値を取るシグモイド関数を設計した。関数の出力値を係数として音韻特徴間で計算した距離に掛け合わせることで修正を行う。

Modification of distance based on sigmoid function

が高いと考えられる。そこで，方向の差が大きい場合に大きな値を取るシグモイド関数を設計した。この関数の出力値を係数として音韻特徴間で計算した距離に掛け合わせることで方向に応じた修正を行う（**図4**）。方向差が小さい場合，係数は1に近い値となるため距離はほとんど修正されず，方向差が大きくなるにつれて大きく修正される。

この方法により，ステレオマイクでは区別できない方向にいる話者どうしの発話に対しては混同を避けつつ，区別可能な方向にいる話者どうしの発話は分類しやすくできると考えられる。

2.4　クラスタリング

クラスタリングには，当社で画像を分類するために開発したkNN Kernel Shift法[3]を応用した。kNN Kernel Shift法は，特徴の密度分布に基づいて分類する方式であり，処理が高速でメモリの使用量が少ないという特長を持つ。

kNN Kernel Shift法によってクラスタを大まかに分類した後，距離の近いクラスタどうしを最短距離法によって逐次的に結合することで最終的なクラスタリング結果を得ている。

3　実験による有効性の検証

開発技術の有効性を検証するため，様々な会議で録音した音声を用いて話者の分類精度を評価する実験を行った。分類精度のよしあしは，話者の分類誤りの指標であるDER（Diarization Error Rate）[4]により評価した。

開発技術との比較のため，二つの従来技術でも評価を行った。一つは音韻特徴間の距離を用いて直接クラスタリングを行う方法であり，もう一つは音韻特徴と方向でそれぞれ算出した距離の重み付き和を用いる方法である。ここで，距離の重みは分類誤りが最小となるよう実験的に調整した。これら二つの従来技術と開発技術の比較を**図5**に示す。

実験には，時間や話者数の異なる九つの会議で録音した音声を用いた。収録した会議の時間は8～172分であり，話者数は3～11人である。ステレオマイクを用いた方向推定では，

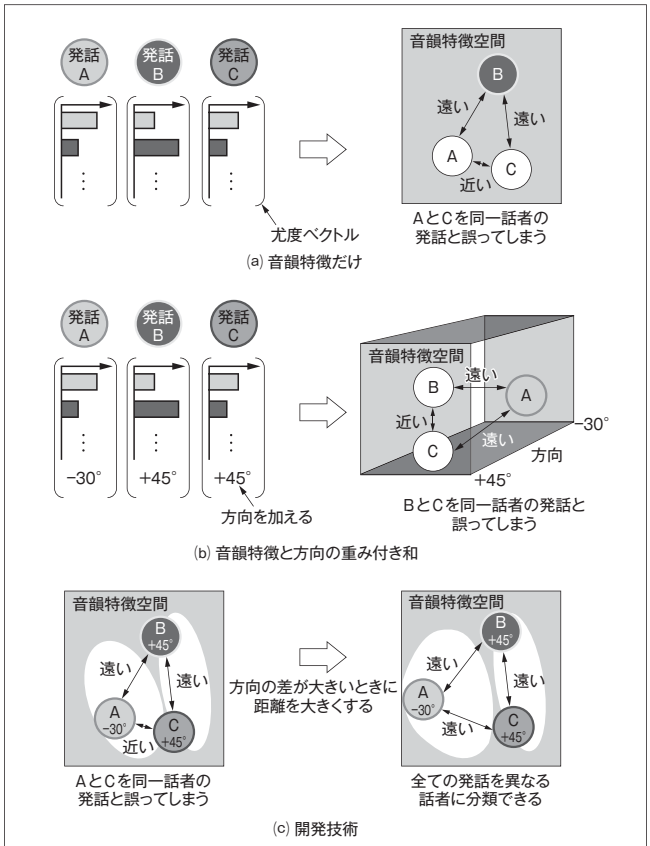

図5．開発技術と従来技術の比較 — 開発技術は，方向の差が大きい場合に限って距離を大きくすることで，区別できない方向にいる話者どうしの発話に対し，混同を避けられる。

Comparison of speaker clustering using conventional and newly developed technologies

同じ方向にいる話者のほか，二つのマイクを結んだ線を対称軸にした鏡像方向にいる話者どうしを区別できない（**図6**）。そこで，九つの会議を，全ての話者の方向がステレオマイクで区別できるかどうかによって二つのグループに分け，グループごとにDERの平均を算出した。

実験結果を**図7**に示す。はじめに，音韻特徴と方向の重み付き和を用いた方法に着目すると，全ての話者の方向が区別できるグループに対しては音韻特徴だけを用いた方法に比べて分類誤りが少なかったが，区別できない方向に話者がいるグループに対しては分類誤りが多かった。これは，音韻特徴と方向の二つの距離による重み付き和を用いたことで，区別できない方向にいる話者どうしの発話において，相対的な距離が近づき混同されやすくなったためだと考えられる。

次に，開発技術に着目すると，二つのグループのどちらにおいても，音韻特徴だけを用いた方法に比べて分類誤りを削減できたことがわかる。このことから，ステレオマイクでは区別できない方向にいる話者どうしの発話に対しては混同を避けつつ，方向を生かして分類精度を高められる開発技術の効果を確認できた。

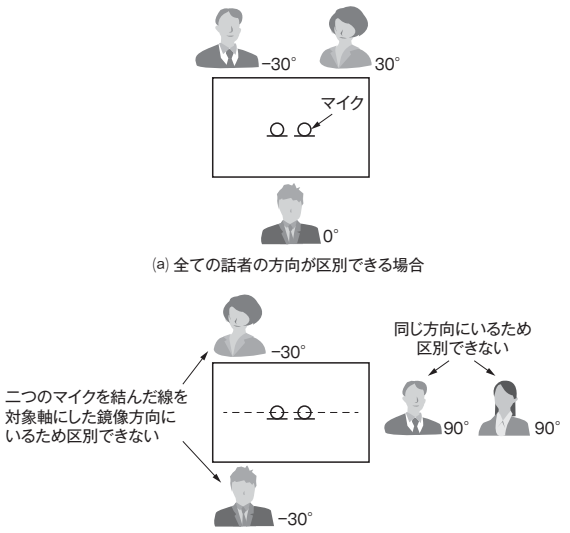

図6. ステレオマイクを用いた方向推定の例 — ステレオマイクでは，話者の位置によっては方向を区別できない場合がある。

Examples of DOA estimation using stereo microphones

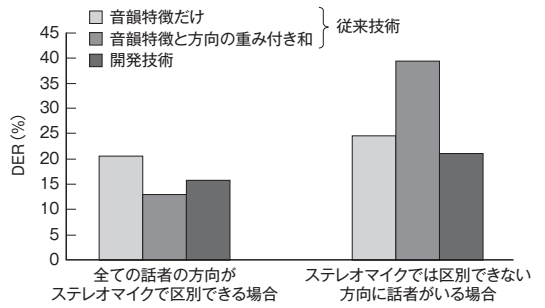

図7. 話者分類実験の結果 — 音韻特徴と方向の重み付き和では，区別できない方向に話者がいる場合に誤りが増加してしまう。一方，開発技術では誤りを削減することができた。

Results of speaker clustering experiment

また，開発技術の処理時間を測定したところ，九つの会議のうちもっとも時間の長かった172分の会議音声に対し，録音が終了してから結果を出力するまでに要した時間は約5秒であった。測定に用いたパソコンのCPUはIntel(†) Core(†) i7-2620M (2.7 GHz) で，搭載メモリは4 Gバイトである。このことから，開発技術は長時間の会議音声でも実用的な時間で処理が行えることを確認できた。

4 あとがき

音韻特徴に加え，ステレオマイクで推定した声の方向情報を必要に応じて併用する新しい話者クラスタリング技術について述べた。開発技術は，推定した方向の差が大きい場合に限って音韻特徴間で計算した距離を大きくすることで，ステレオマイクでは区別できない方向にいる話者どうしの発話に対しては混同を避けつつ，方向を生かして話者の分類精度を高められる特長がある。様々な会議で録音した音声を用いた話者の分類実験により，開発技術は従来の技術より高い精度で話者を分類でき，更に，長時間の会議音声でも実用的な時間で処理が行えることを確認できた。

今後は，話者の分類性能の更なる改善を進めるとともに，会議の効率的な振返りを支援する新たな差異化機能を実現する技術を開発していく。

文 献

(1) Anguera, X. et al. "Automatic weighting for the combination of TDOA and acoustic features in speaker diarization for meetings". Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. Honolulu, HI, USA, 2007-04, IEEE Signal Processing Society. IEEE, 2007, p.241－244.
(2) 丁 寧 他. “音韻情報と方向情報を用いた発話間距離による話者クラスタリング”. 日本音響学会2014年秋季研究発表会講演論文集. 札幌, 2014-09, 日本音響学会. 2014, 論文番号2-Q-8.
(3) Hirohata, M. et al. "KNN Kernel Shift Clustering with Highly Effective Memory Usage". Proc. The Twelfth IAPR Conference on Machine Vision Applications. Nara, Japan, 2011-06, MVA Organization. 2011, p.393－396.
(4) Ishiguro, K. et al. Probabilistic Speaker Diarization With Bag-of-Words Representations of Speaker Angle Information. IEEE Trans. Acoustics, Speech and Signal Processing. **20**, 2, 2012, p.447－460.


木田 祐介 KIDA Yusuke
研究開発統括部 研究開発センター 知識メディアラボラトリー 研究主務。音声信号処理及び音声認識の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.


丁 寧 DING Ning
研究開発統括部 研究開発センター 知識メディアラボラトリー。音声信号処理の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.


広畑 誠 HIROHATA Makoto
研究開発統括部 研究開発センター 知識メディアラボラトリー 研究主務。音声信号処理の研究・開発に従事。日本音響学会会員。
Knowledge Media Lab.

一般論文