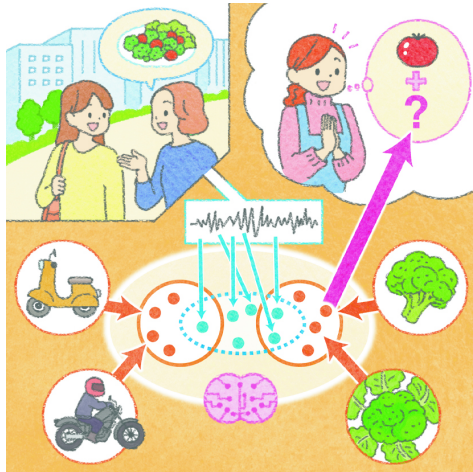




背景

複数の話者の声や物音などが混在した音響信号から目的の話者の声を分離して取り出す信号処理は音源分離と呼ばれ、音声インタフェースにおいて重要な役割を果たしています。従来、音源分離では、音の方向、声質、声の高さなど、物理的な手掛かりに基づく方法が用いられていました。

成果の概要

これに対し本技術では、長時間の音響信号に対して、意味内容（概念）を手掛かりにして、目的とする話題を話している話し声がどの時点に存在するかを特定し、さらにその目的の音声だけを分けて取り出す処理を世界で初めて実現しました。

ConceptBeam は意味に基づく新しい音源分離技術です。

入力信号に、同時に複数の音（複数話者の音声）が混在していても、画像や音声などで指定した関心のある話題の音声を検出し、その関心のある音声だけを分離・抽出することができます。

入力

複数の話者・話題が混在した音声



目的指定

画像や音声で指定



出力

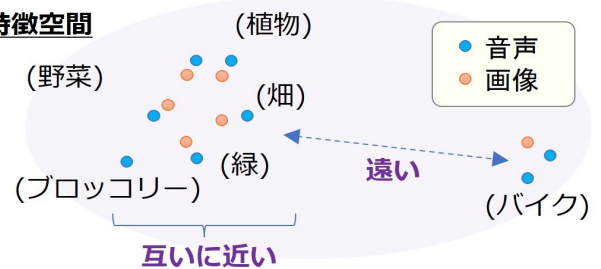
該当箇所の目的音声を分離・抽出



技術のポイント

私たちが開発した機械学習と信号処理の技術では、音声や画像といったデータの種類に関わらず、信号の断片どうしの意味的な近さを測定できます。これにより、入力信号に含まれる、指定された画像や音声に意味的に近い成分を特定し、分けて取り出すことが可能になりました。

特徴空間



技術のポイント

- 入力信号と目的情報とを概念空間でマッチングすることで、複数話者の声や物音などが混在した音響信号に対しては精度が低下しがちな音声認識によることなく、情報の検索や分離抽出が可能
- 目的情報の指定（概念指定）は、音声、画像、テキストのいずれか、またはそれらの複数の組み合わせにより可能
- 言語や情報の種類をまたがった概念指定（例：英語の入力音声に対して、日本語の音声や画像で目的概念を指定）をサポート

この研究がもたらす未来

ネット上や実世界での膨大なメディア情報の中から、言語情報・非言語情報によらず、また音声や画像といった情報の種類に関わらず、必要な情報をよりの確に選び出し活用できるようになります。

出展企業

日本電信電話株式会社

問い合わせ先

rdforum-exhibition@ml.ntt.com