

非負値行列因子分解を用いた雑音環境下での音声認識法に関する研究

杜 雨軒 (1210033)

北陸先端科学技術大学院大学 情報科学研究科

2014 年 2 月 12 日

キーワード: 雑音環境, 音声認識, 聴覚情景解析, 非負値行列因子分解.

音声認識は音声科学における重要な分野の一つである。音声認識に関する研究は、1950 年代から続いている。自動電話応答や、音声認識カーナビや、Speech-to-text の分野などに対し、音声認識が重要な役割を担ってきている。現実の生活で、音声認識の応用に伴い、手が離せない状態などの状況下で簡単に音声でコンピュータを操作することが可能である。現有の音声認識技術は雑音のない状況下で、優れた認識精度を得ることが可能である。しかし、現実の生活では、常に雑音が存在する。雑音が音声認識システムの入力に混在すると、システムの認識精度が著しく劣化する。このため、現時点では雑音環境下で音声認識の応用は困難である。

雑音の影響を受けやすい問題に対して、過去の研究としては、雑音除去を前処理として認識モデルに付加したり、音響モデルを変形し雑音に適応させる手法がある。しかし、いずれの手法でもその局限性があり、この問題を解決することが困難である。一方で、人間は雑音と目的音が混在する状況でも聴取できる聴覚能力を持っている。この聴覚能力を考慮し、羽二生らが「聞き耳」モデルを提案した。「聞き耳」モデルは、ある目的音が雑音音声に存在すると仮定する。仮説から、その目的音の知識を既知情報として扱い、雑音と目的音を分離する。妥当に雑音と目的音を分離できれば、仮説の中の目的音が雑音音声に存在するといえる。「聞き耳」モデルは、このような分離の妥当性により、目的音を認識する「仮説・検証」型の音声認識法である。この方法は、人間の聴覚処理をまねするため、雑音への頑健性を示した。しかし、この方法の処理プロセスには、膨大な計算量が必要であり、実用的な手法ではない。

「仮説・検証」型の音声認識法のコンセプトが有効であると考え、本論文では、このコンセプトを実現するために、新たな方法を用い音声認識手法を提案する。また、本提案法の有効性について検討する。

音声認識のためのテンプレートを作成するため、テンプレートを入力に合わせて微調整できるテンプレート合成法として、本研究では Modified Restricted Temporal Decomposition

(MRTD) 手法を用いた。また、認識用テンプレートを用いて、目的音と雑音を分離できる音声分離法として、本研究では処理の高速化が期待できる Non-negative Matrix Factorization (NMF) 手法を用いた。この 2 つのツールを用い、本研究で音声認識法を構築した。

背景音として、white、pink、babble、factory の 4 種類雑音が存在する状況で、1 名話者が日本語 4 モーラ単語を 100 個発話する状況を想定した。この状況下で、提案法における音声認識モデルを実装し、音声認識のシミュレーションを行った。SNR は、0 dB、10 dB、20 dB、 ∞ と設定した。比較するため、本研究では Dynamic Time Warp (DTW) における音声認識法を用い、同じ雑音環境下で比較実験を行った。シミュレーションの結果から、すべての雑音環境に対して提案法における音声認識法は、DTW における音声認識法より常に正確な認識結果を得られることが分かった。特に、0 dB の環境下で提案法における音声認識法の認識率は、DTW における音声認識法の認識率より 50% 以上高いことが分かった。さらに、提案法における音声認識法が目的音を認識する所用時間は、「聞き耳」モデルの所用時間より圧倒的に短いことが分かった。

以上のことから、4 種類の雑音と日本語単語が加算された状況において、提案法が有効であることが確認できた。また、期待する通りに、「仮説・検証」型の音声認識のコンセプトを実現する上で、提案法の処理時間が「聞き耳」モデルより、大幅に短縮できた。