修 士 論 文

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

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

杜　雨軒

2014 年 3 月

修 士 論 文

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

指導教員　赤木正人　教授

審査委員主査　赤木正人　教授
審査委員　鵜木祐史 准教授
審査委員　党建武　教授

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1210033杜　雨軒

提出年月: 2014 年 2 月

概 要

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。

目 次

図 目 次

表 目 次

第1章　序論

1.1　はじめに

音声認識 (Speech Recognition) とは、人の話す音声言語をコンピュータにより解析し、話している内容を文字データとして取り出す処理のことである [1]。キーボードやマウス、タッチパネルの代わりに、音声を使い音声認識技術を生かすことで、コンピュータを操作することが可能である。例えば、手に障害のある人や、手が離せない状態で運転する場合に、音声で便利に操作できる Siri や音声認識ナビなど、音声認識は生活の中で広く使われている。このような音声認識技術に関する研究は 1950 年代から行われている [3]。現在までに行われている音声認識手法のとしては、DP (Dynamic Programming) マッチング法、HMM (Hidden Markov Model)、ニューラルネットワークを用いた手法が代表的なものである。しかし、上記の手法は、雑音が入力音声に混在しない、理想的な状況でしか有効ではない。現実生活には、常に雑音が存在するので、音声認識が雑音環境に弱い問題は、まだ完全に解決できない。音声認識を実用化するため、雑音環境下や実環境で認識率を向上させることが次の課題となっている。

1.2　研究背景

1.2.1　音声認識における問題点

音声認識システムの概要を、図 1.1 [2] に示す。特徴分析とデコーダにより、入力音声が単語列に変換される。デコーダには、必要な情報源として音響モデル (Acoustic Model) 、辞書 (Dictionary) 、言語モデル (Language Model) がある。これらの情報源に、音響モデルはある音素がどのような特徴量系列に対応するのかの情報を持っている。言語モデルはある言語で単語がどのように並ぶのかについての情報を持っている。辞書はある単語の発音に対応する音素の並びの情報を持っている。これらの 3 つの情報を用い、考えられるすべての単語の並びについて、その単語列と入力の特徴系列と対応する尤度を計算し、最も尤度の高い単語の並びを認識結果とする。

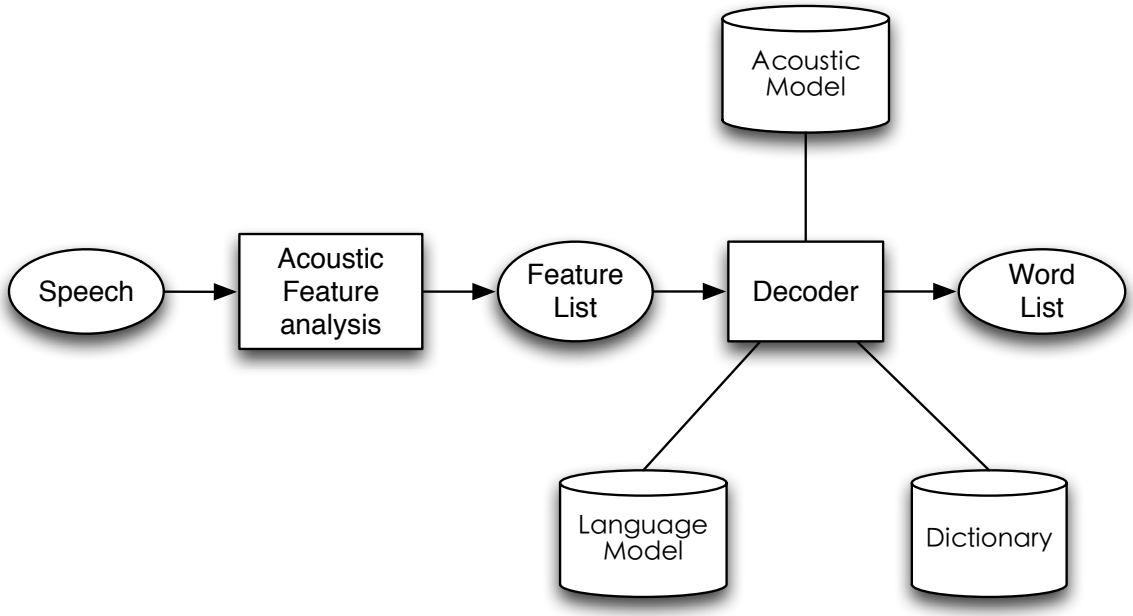


図 1.1: 音声認識システムの概要

このプロセスで、もし雑音が入力音声に混在すると、特徴分析で得られた特徴量は雑音の影響を受け、変化を生じる。音響モデルの持つ情報はクリーンな環境下で用意した特徴量であるので、雑音音声の特徴量との対応は間違いやすい。この問題を説明するため、図 1.2 [3] が示したのは、同じ音声の特徴量がクリーン環境下と SNR (signal-to-noise ratio) が 10 dB の雑音環境下での差である。この中に、示された特徴量は、音声認識システムでよく使われる、1 次のケプストラム係数という特徴量である。図から雑音音声とクリーン音声の差が大きくあることが分かる。このため、音声認識が雑音に弱い問題の根本は、音響モデルが持つクリーンな環境下で抽出された音響特徴量（認識レファレンス）と雑音環境下での音響特徴量が大きく異なっているため、尤度を計算することより誤認識しやすいことである。

また、図 1.3 に示したのは雑音環境下で劣化した音声認識の認識率である [4]。認識されたデータは一話者が発話した、1 から 10 までの 10 個の英語数字である。音響モデルで、クリーンなデータが用いられた。入力データのクリーン音声に SNR が 0 dB から 20 dB の 3 セット雑音をたし、雑音環境をシミュレートした。この 3 セットの雑音では、それぞれ set A (subway, babble, car, exhibition). set B (restaurant, street, ariport, station), set C (subway, street) の雑音である。結果としては、もともとクリーンな環境下で 100%近く認識できる音声認識法が 3 セットの雑音環境下で、認識率は大幅下回った。さらに、話者の変化または認識データ数の増加とともに、音声認識率がさらに下がる可能性もある。

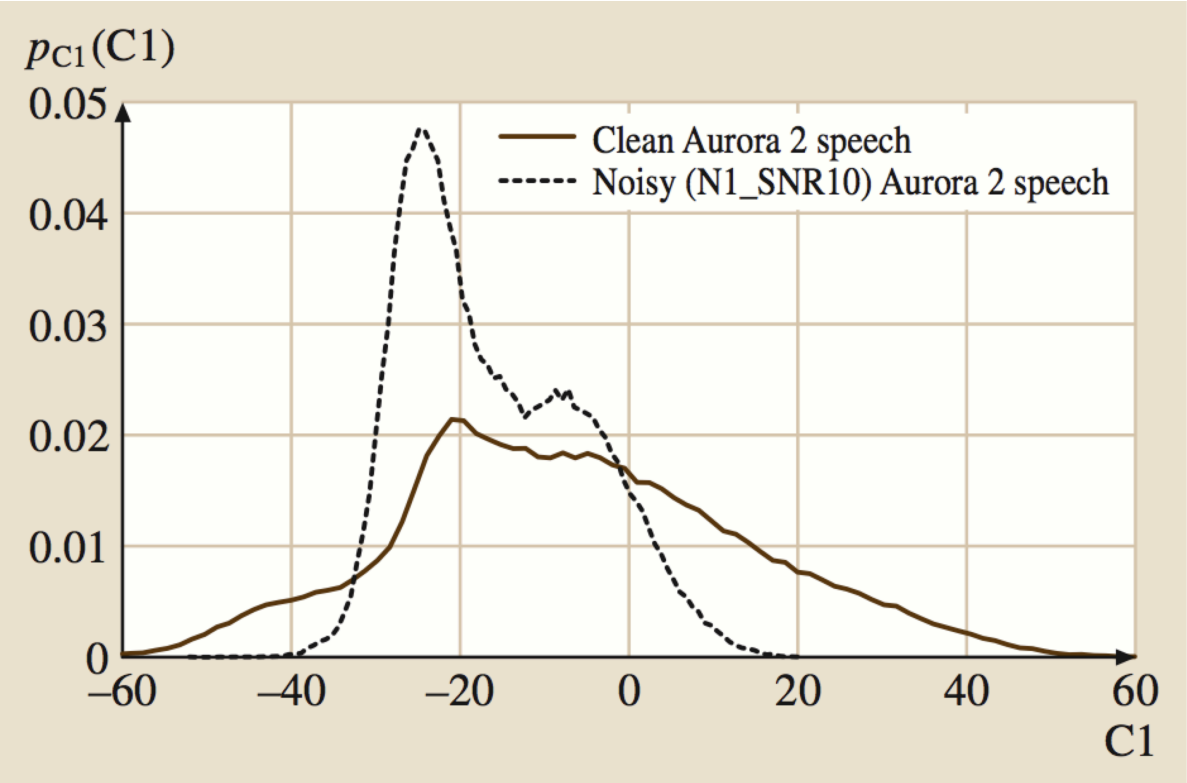


図 1.2: クリーンと雑音環境下での特徴量

Dautrich らの研究 [5] によると、音声認識システムは雑音対策をしない場合に、SNR が 24 dB（30 dB はかなりクリーン）以下であると、音声認識率の劣化が始まる。このため、実環境に頑健な音声認識システムを構築するため、雑音の影響を抑える方法が必要である。

雑音環境下での音声認識率を向上させるため、研究されている方法は大別にして 2 種類がある。一つ目は、デコーダに雑音音声を入力する前に、雑音除去手法を前処理として用いる。これは雑音除去より、デコーダにクリーンな音声を入力する前処理による雑音環境への対応の考え方である。二つ目は、入力の雑音音声をそのままデコーダに入力する。続いて、音響モデルが持つ認識レファレンスを機械学習などの方法で雑音環境に適応させ、認識レファレンスと雑音音声の特徴量を近づけることができる。この考えは、音響モデルの変形による雑音環境への対応手法である。

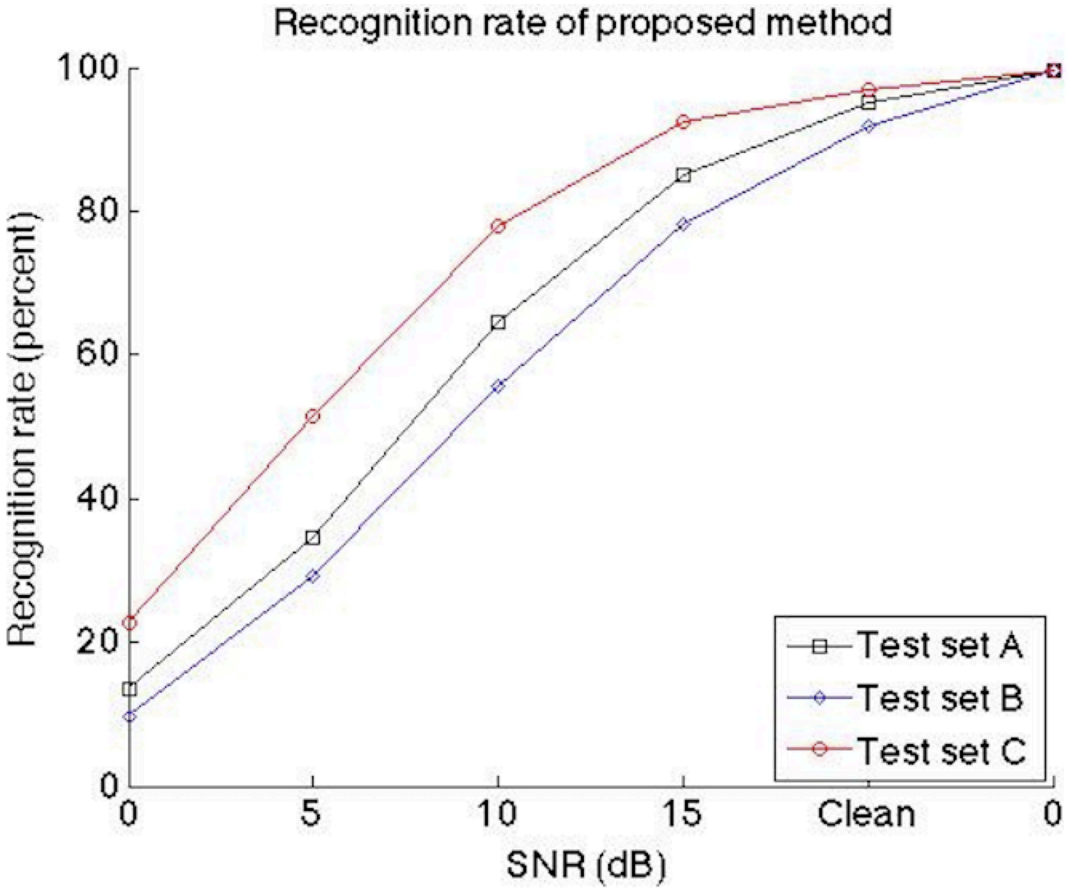


図 1.3: 雑音の影響を受けた音声認識率

1.2.2 前処理による雑音環境への対応

既存の認識手法に雑音除去のような前処理を加えることにより、雑音環境に適応する手法がいろいろ研究されている。これらの手法は、雑音除去プロセスにマイクロホンを 1 つ用いる手法とそれ以上用いる手法に大別される。1 つのマイクロホンを用いる手法は、周波数領域における信号処理が中心である。2 つ以上のマイクロホンを用いる手法は時間領域での処理が中心となる。

1 つのマイクロホンを用いる代表的な手法としては、スペクトルサブトラクション法 (SS) [6] がある。この手法では、時刻 t における観測信号 $y(t)$ は音声信号 $s(t)$ と雑音信号 $n(t)$ の線形和で表現できると仮定する。

$$y(t) = s(t) + n(t) \tag{1.1}$$

$s(t)$ と $n(t)$ が独立であれば、フーリエ変換より、式 (1.1) は

$$
\begin{aligned}
Y(f) &= S(f) + N(f) \\
S(f) &= Y(f) - N(f)
\end{aligned} \tag{1.2}
$$

となる。よって、周波数領域において、$Y(f)$ から $N(f)$ を引き去ることより、クリーンな音声の周波数特徴量が得られる。正確な $N(f)$ を知ることは困難であるが、$S(f)$ のないときに $\hat{N}(f)$ を推定することができる。

マイクロホンを2つ以上用いるマイクロホンアレーによる雑音抑圧 [7] は、マイクロホンアレイ上の各マイクロホンから得られる信号から、特定の方向だけの音を抽出する。ビームフォーミングで多く用いられるのは、音源と各マイクロホンからの距離の差により時間差を利用した遅延和アレイである。これらの情報を利用し、観測信号から各音源で生じた音が推定できる。

しかし、スペクトルサブトラクションは定常雑音を想定しているため、非定常雑音や突発雑音が入力音声に混在する場合には有効ではない。また、マイクロホンアレーにおける雑音除去手法は、マイクロホンの数は音源の数と同数またはそれ以上であることが必要である。実環境では、音源の数は未知であり、マイクロホンアレーにおける雑音除去の音声認識への応用は難しい。

1.2.3 音響モデルの変形による雑音環境への対応

雑音のスペクトルは常に変化している。このため、雑音の特徴量も統計的手法であるマルコフモデル (HMM) で表現し、さらに無雑音音声の HMM と雑音の HMM から、目的の雑音環境の音声 HMM 合成ができる。この方法は、PMC (Parallel Model Combination) [8] または NOVO [9] と呼ばれている。この方法では、学習用の雑音付加音声は必要なく、雑音モデルだけを用意すればよいという利点がある。

しかし、HMM 合成法は比較的計算量が多いため、非定常雑音環境に対応できるが、短時間にモデルを再合成することが難しい。さらに、雑音の性質が突然変化したり、突発的な雑音やモデル化されていない雑音が生じると、認識率が下がるという問題点がある。

1.2.4 聴覚情景分析による雑音環境への対応

機械は音声を認識するときに雑音に影響されやすい。一方で、人間は劣化している音声または雑音下での音声にたいして、非常に頑健性を持っている。人間は耳から入る物理量（音響特徴量）だけではなく、脳からの情報 (トップダウン情報) を積極的に利用し、聴取できる [10]。現有の音声認識システムの言語モデルや音響モデルでは、

- Parsing: 信号処理領域でトップダウン情報を記述

- N-gram, N-phone, HMM: 抽出された音響特徴のつながりを確率的に記述することにより、音声情報を数十 ms の範囲内で記述

と考えられるが、トップダウン情報は時間的にはもっと広範囲にわたる。しかも、これらの手法では、音響特徴の抽出・再構築まで及ぶトップダウン情報は見受けられない。

人間は二つ以上のメセッジが混在していても一方を選択的に聴取することが可能である。このような聴覚上の効果は「カクテルパーティ効果」と呼ばれている [11]。例えば、3 人が「おはよう」、「こんにちは」および「こんばんは」を同時に発話する。聞き手は事前情報を持っていない場合に、3 人の言葉が混在しているので、聴取が難しい。しかし、もし事前に聞き手に次の発話に「おはよう」があるという情報をあらかじめ伝えれば、聞き手は簡単に「おはよう」を聞き取れる。人間は常にこのような情報を利用し、聴取できる優れた聴覚能力を持っている。この能力を音声認識技術に応用できれば、雑音に頑健な音声認識の構築も可能である。

カクテルパーティ効果の重要な要素として、聴覚的な「情景解析」 (scene analysis) がある。人間の聴覚情報処理プロセスには、周囲のすべての音声が重畳された状態となった音声をいったん部品に分解する。分解した後に、強く関係する部品同士をまとめ、周囲の状況を把握する。この聴覚の一連の情報処理過程は「聴覚情景解析」Auditory Scene Analysis (ASA) と呼ばれている [12]。Bregman によれば、人の耳に届いた音は、部品に分解され後に、各音源から生じた音声の一連の部品同士をグルーピングを行い ASA を行っている。この現象は多くの心理実験に基づいて報告されている [12]。人間は意図的に目的音に対して注意を向け、混在した音声を部品に分解し、Bregman の法則を用い、グルーピングで聞き取りたい部品同士をまとめる。この行為は「聞き耳」と呼ばれている [10]。

「聞き耳」のコンセプトを用い、羽二生らが「聞き耳」の能力を模擬するモデルを提案した [13] [14]。この「聞き耳」モデルの概要は図 1.4 で示されている。認識モデルの入力が雑音音声の X_N である。この X_N の中に、あるターゲット音声 v が存在する。X_N に存在しうる v の候補は $v = 1,\ 2\ ,\ \ldots\ ,\ V$ （この数値はターゲット音声の ID）である。最初に $v = 1$ と仮定する。すなわち、$v = 1$ の対応するテンプレート C_1 が X_N の特徴量の中に存在すると仮定する。続いて、C_1 をもちいて、X_N の特徴量の中から、二波形分離法 [15] で C_v と雑音を分離する。この二波形分離法は、入力音を部品にいったん分解し、Bregman の法則 [12] により部品同士をまとめ、再合成する音声分離手法である。さらに、「聞き耳」モデルには C_1 が X_N に存在するという仮説があるため、音声を分離する際に X_N をいったん音声の部品に分解し、C_1 の部品同士とそれ以外の部品同士をグルーピングし、再合成する。このプロセスは図 1.4 の中の Segregation と対応する。Bregman の法則により、下記の 2 つの状況から仮説の妥当性を判断できる。

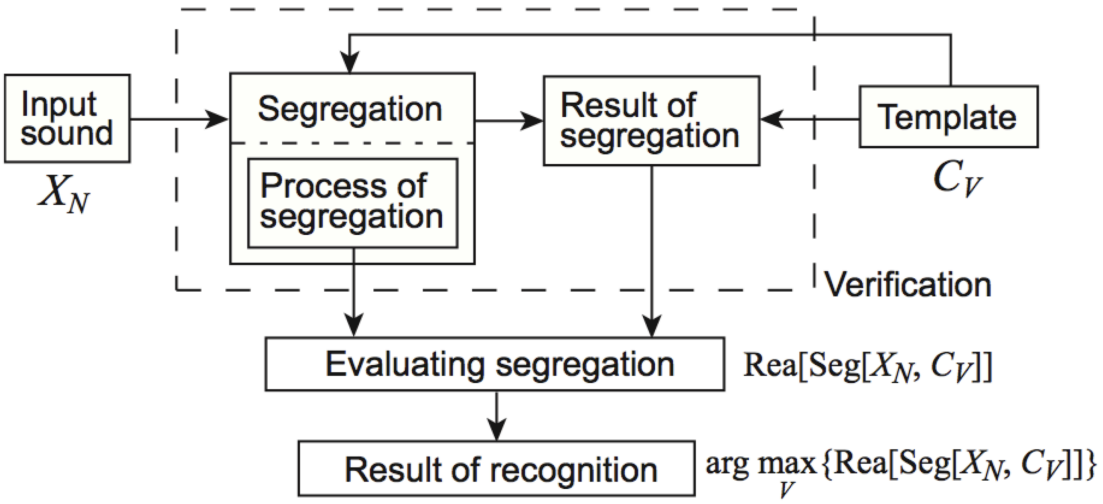


図 1.4: 聞き耳モデルの概要

- C_1 が X_N に存在する仮説が妥当ではない時

  X_N で分解した部品の中に C_1 に対応する部品が存在しない場合がある。この時 Segregation を中止し、$v = 1$ が雑音音声 X_N に存在しないと判断する。また、もしある v が X_N に存在し、C_v の部品の一部分は C_1 の部品とグルーピングできるため、Segregation プロセスが中止せずに処理する。しかし、Segregation が最後まで処理できても、その分離の結果から C_1 に対応する部品同士を再合成し、合成した音声は $v = 1$ ではないはずである。すなわち、$v = 1$ が X_N に存在する可能性が低いと判断する。

- C_1 が X_N に存在する仮説が妥当である時

  Segregation が最後まで処理でき、分離の結果から C_1 に対応する部品同士を再合成し、合成した音声は $v = 1$ となるはずである。

このように、Segregation のプロセスと結果を評価し、$v = 1$ が X_N に存在する可能性を計算することは、図 1.4 の $Rea[Seg[X_N, C_V]$ と対応する。上記の分離・検証プロセスは $v = 1$ だけではなく、すべての $v = 1,\ 2\ ,\ ...\ ,\ V$ の候補に対して X_N を分離し、X_N に存在する可能性を計算する。認識結果は X_N に存在する可能性の最も高い v である。すなわち、認識結果 v が図 1.4 の $arg\ max_V\{Rea[Seg[X_N,\ C_V]]\}$ である。

この方法は人間の聴覚能力を模擬していて、前処理や雑音のモデル化などがいっさい不要である。全部 6 種類の雑音にたいして、羽二生らが日本語数字の認識実験を行った

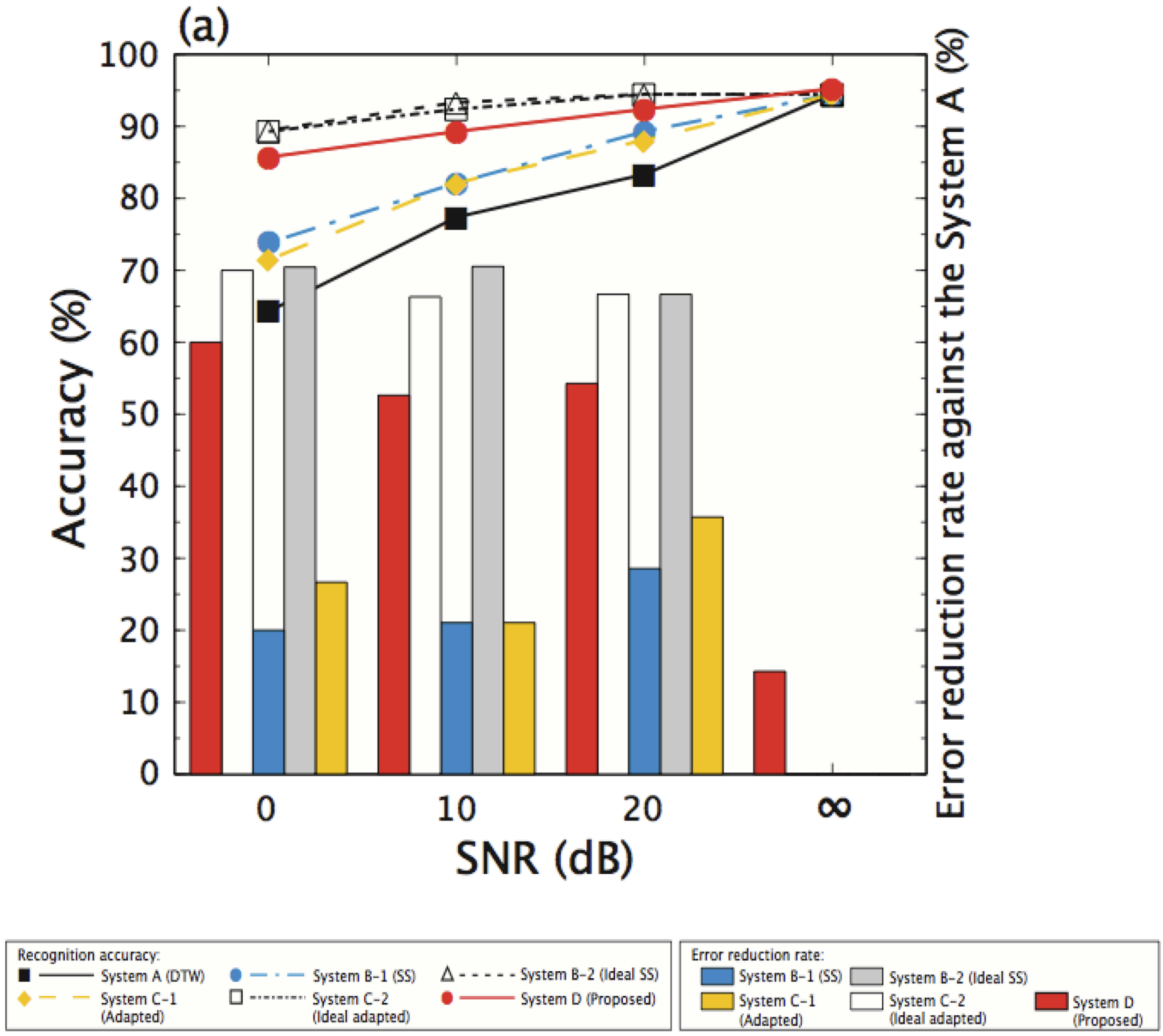


図 1.5: 聞き耳モデルの認識結果

[14]。6 種類の雑音はそれぞれ、(a) machine gun noise、 (b) babble noise、 (c) pink noise、(d) destroyer operations room background noise、 (e) military vehicle noise と (f) white noise である。雑音環境を 0 dB からクリーンまで設定した。「聞き耳」モデルと比較した ASR システムは：

- システム A：前処理または音響モデルの変形なし
- システム B-1：SS 法の前処理（雑音未知）
- システム B-2：理想的な SS 法の前処理（雑音既知）
- システム C-1：音響モデルの変形における雑音適応法（雑音未知）

- システム C-2：理想的な音響モデルの変形における雑音適応法（雑音既知）
- システム D: 「聞き耳」モデルにおける音声認識法

となる。「聞き耳」モデルにおける Automatic Speech Recognition (ASR) システムの (a) 雑音環境下での認識結果を、図 1.5 に示した。結果から、「聞き耳モデル」は雑音情報が未知情報として扱っているが、雑音が未知の音声認識システム（システム B-1 とシステム C-1）と比べてよい認識率を得た。また、雑音が既知の理想的な音声認識モデル（システム B-2 とシステム C-2）とほぼ同じ程度の認識結果を得た。すなわち、「聞き耳」モデルが前処理また音響モデルの雑音適応していない音声認識法でありながら、多種類の雑音環境に頑健であることが分かった。

しかし、この方法は二波形分離モデル [15] を Segregation 法として用い、音声を部品に分解し、グルーピングそして再合成している。このことには非常に時間がかかる問題点がある。例えば、「聞き耳」モデルにおける音声認識法は数字を認識するときに、候補とする v が 10 個しかない場合でも、1 つのデータを認識するため 1 日以上かかる。この方法を拡張すれば、候補を増加する必要があり、さらに時間がかかる恐れがある。上記の原因で「聞き耳」モデル [13] [14] が優れた頑健性を持っているにもかかわらず、実環境での応用は難しい。

1.3 研究目的

既存の手法において実環境に応用できる音声認識システムはない。羽二生らの方法では、ASA のコンセプトを用い、まずある目的音候補 v が雑音音声 X_N に存在すると仮定する。この仮説により、目的音候補 v_h の情報を用い、雑音と目的音を分離する。すべての目的音候補の情報を用い、雑音と目的音を分離したあと、分離の妥当性を評価する。評価の結果により、目的音を認識する方法は、「仮説・検証」型の音声認識手法である。この「仮説・検証」型の音声認識法は、非常に雑音への頑健性を示した。

そこで、本研究の目的を、人間の聴覚能力を考慮したコンセプト（ASA）に基づき、雑音に頑健かつ実用化可能の「仮説・検証型」の音声認識手法を提案することとする。

1.3.1 問題設定

音声認識という概念は、応用される状況により問題の設定が大きく変化する。この変化により、音声認識システムは極めて複雑になる可能性がある。このため、音声認識システムを構築する前に、問題設定の範囲を決める必要がある [3]。代表的な問題を以下にあげる。

- 認識するユニット：システムが音声を認識する際に、処理する最小のユニットである。このユニットの範囲は単語、音節、音素である。不連続語彙を認識する際、音

素に比較すると単語認識がシンプルである。一方で、連続語彙を認識する際に、単語をユニットとして処理することは困難である。

- 認識できる語彙のサイズ：語彙サイズは音声認識システムが認識できる語彙の数から、小型（2-100 語彙）、中型（100-1000 語彙）、大型（1000 語彙以上）と分けられている。語彙サイズが小さければ小さいほど、認識処理がシンプルである。

- 発話モード：音声認識システムの入力を表す問題設定である。設定の範囲は単独語彙、連続語彙（例えば、連続の数字の認識）、連続発話となっている。単独語彙を認識するでは、認識タスクはそれぞれ独立であるが、連続語彙の認識には各語彙の関連性を考える必要がある。さらに、連続発話の認識には、正しく認識するため文法などの考慮も必要となる。

- 発話者モード：人間の発話が同じ言葉を複数回発話しても、特徴量が一定ではない。さらに、異なる発話者の個人性により、音声認識システムに影響を与える可能性が高い。発話者モードは、特定話者（指定された話者の発話を学習させる。入力はこの指定話者の発話のみである）、話者適応（入力の発話者に対する適応ができる）、話者独立（適応させずに、発話者の発話が認識できる）と分けている。

- 発話環境：発話者の発話環境によっては、雑音などの影響を受け、音声認識システムの認識率は大幅に劣化する。また雑音が入力音声の中に入れば入るほど、音声認識の処理が困難となる。この雑音への頑健性については、まだ完全に解決できないため、多数の方法が研究されている。

前処理による雑音環境への対応や音響モデルの変形による雑音環境への対応を行う音声認識システムは、前節で述べた問題点より実環境に対応することが困難である。しかし、人間の優れた聴覚能力を考慮し、聴覚情景解析のコンセプトを利用した例としては、「聞き耳」モデル [13] [14] が雑音への頑健性を示した。一方で、この方法は計算量が膨大で、実用化することが困難である。本研究では、上記の問題設定に従い、「聞き耳」モデルと同じなコアコンセプトに基づき、より高速化が期待できる新しい方法を用い、認識モデルを構築する。最終的な目標は、提案法における音声認識システムが実環境下で音声認識を行うことであるが、本論文では「仮説・検証」型の音声認識法のコンセプトを実現する新たな方法を用い、提案法の有効性を検討する。

すなわち、本研究の目的は ASA のコンセプトに基づき、新たな手法で雑音と目的音の分離の妥当性を判断し、「仮説・検証型」の音声認識モデルを構築することである。最終の目的は実用化向けの音声認識システムを構築することだが、本論文では完全または複雑な音声認識システムを構築することではなく、新しく提案する方法の有効性の検証を試みる。このため、本研究の問題設定は雑音の影響の問題に着目し、ほかの問題を簡略化する。本論文では、音声認識モデルの認識ユニットは単語であり、語彙サイズは小型である。また、発話モードは単独語彙であり、発話者は特定話者と想定している。

1.4　本論文の構成

本論文は 5 章で構成する、本章では音声認識の背景、問題点また本研究の目的を述べる。各章の概要を以下に示す。

- **第 2 章 提案法の概要**

第 2 章では、提案法の概要および全体図を述べる。また本研究で用いられる手法の概要を記述する。

- **第 3 章　提案法の実装**

第 3 章では、まず本研究で用いられる各手法の詳細を説明する。続いて、各手法は本研究での使い方を述べる。

- **第 4 章　評価実験**

第 4 章では、本研究の有効性を評価するため、評価実験を行う。評価実験のデータや条件などを述べ、実験結果を示す。また、結果にたいして考察を行う。

- **第 5 章　結論**

第 5 章では、得られた結果をまとめ、今後の課題を述べる。

第2章　提案法の概要

本章では、本研究の全体的な概要を記述する。また、音声認識法を構築するため、本研究でツールとして用いる手法を説明する。

2.1　提案法の概要

前節の問題設定に従い、本研究の概要を図2.1に示す。入力音声 (Input sound) X_N の中に含まれる v を認識するため、まず v_h が X_N に存在すると仮定 (Hypothesis v_h) する。仮説により、テンプレート C_v (Template C_v) の情報を用い、雑音と目的音を分離 (Separation) する。分離の結果 (Result of separation) の妥当性を検証する (Evaluating separation result) ことにより、目的音を認識する。本研究の認識モデルは、図2.1で示したような「仮説・検証」型の音声認識モデルである。

提案法のモデルで認識する前に、準備段階として認識語彙のデータベースを用意する。システムが認識できる語彙を元々の単語データ (Speech Data) を用いて合成器で表現する。これにより、表現されたデータ (Synthesized Data) が合成器でコントロールできるようになる。すなわち、表現した音声が複数発話で生じる発話データに近づけることができ、音声認識のテンプレートとして用いることが可能となる。さらに、本論文では単独語彙を認識するが、合成器の入力を変えることにより、単語から音素までを提案法のテンプレートにできる。これにより、合成器をもちいて、本研究を音素認識まで拡張することが可能になる。

認識処理を行う際に、認識モデルに入力される雑音音声 (Input sound) は X_N である。X_N では雑音が未知であり、この雑音音声の中にある目的音 v が含まれている。本研究では、話者1名の音声に雑音が重畳している状況であり、X_N に1つだけの v が存在する状況のみを想定している。X_N に存在する v はどの単語かわからないが、言語知識や確率モデルの情報から、X_N に存在しうるすべての v の候補 (Candidates) は $v = 1,\ 2\ ,\ ...\ ,\ V$ となる。本論文の中心は雑音に対応する対策であり、候補の選択法を考えずにすべての認識できる語彙が候補として用いられる。ここで、HMM確率モデルなど (Statistic Model) の方法をもちいて候補をうまく選択できれば、処理の簡略化また高速化が期待できる。すべての候補の中に、ある音声 v_h が X_N に存在する、すなわち目的音が $v = v_h$ と仮定する。この仮説 (Hypothesis v_h) を検証するため、v_h に対応するテンプレート (Template C_v) を用い、X_N を目的音 v_h と雑音に分離 (Separation) する。分離の結果 (Result of separation)

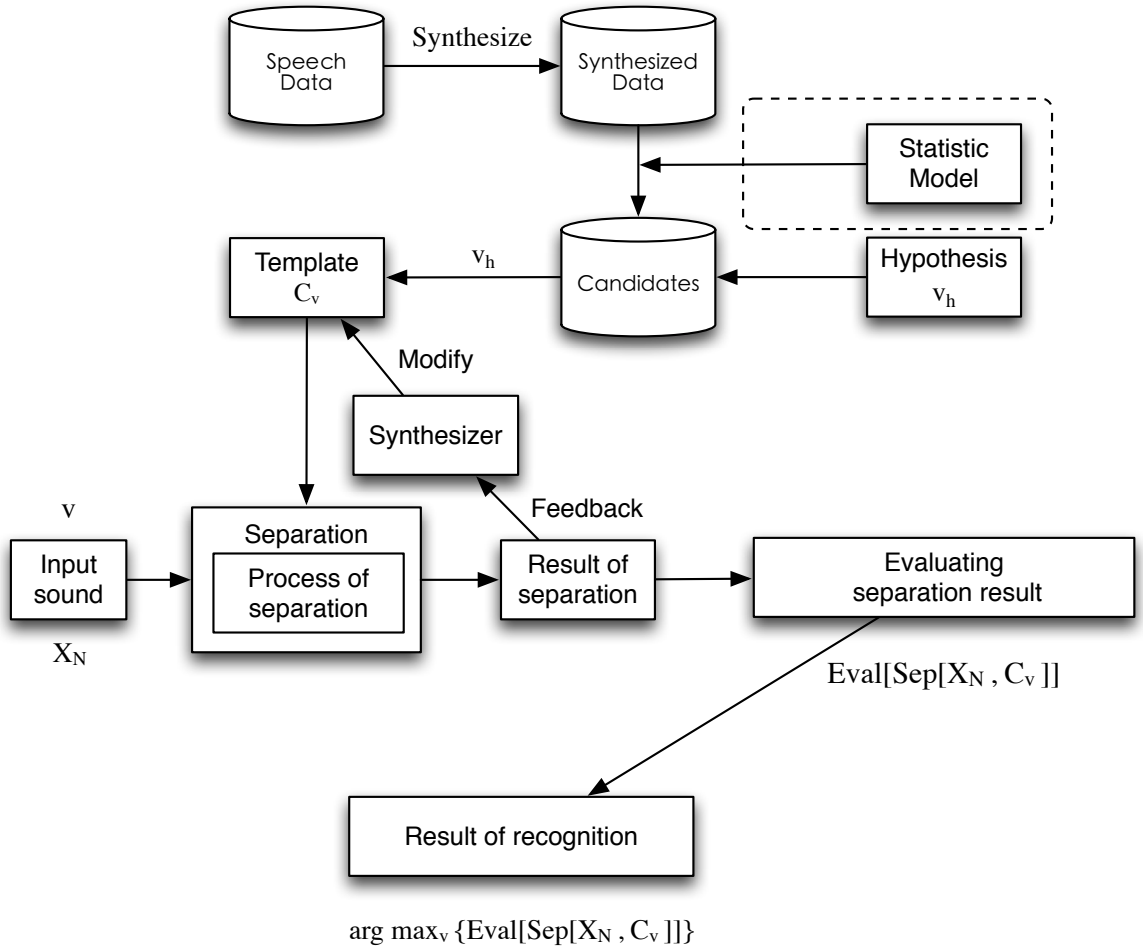


図 2.1: 本研究の概要

を評価すること (Evaluatiing separation result) により、この仮説の妥当性を判断できる。すなわち、仮説・検証のプロセスで音声を認識することができる。

v_h が X_N に存在するという仮説に基づき、v_h と雑音を分離する際に、対応するテンプレート C_v を既知情報としてあつかう。X_N に含まれるのは、ある目的音 v と雑音だけなので、雑音が未知であっても、C_v をもちいると、目的音と雑音を簡単に分離できる。また仮説の妥当性、すなわち v_h が本当に X_N に含まれるかどうかにより分離の結果が大きく異なる。この異なりがどの候補が X_N に存在するかの判断基準となる。

また、複数回の発話で音声の時間域また周波数域の変化を吸収するため、目的音と雑音を分離した結果を用いて、合成器にフィードバックを与える。このフィードバックを用いて入力音声に近づけるため、テンプレートを微修正することができる。修正により、正しい候補のテンプレートを用いる分離結果と誤った候補の分離結果のばらつきが大きくなり、認識率の向上が期待できる。十分修正したテンプレートを用い、目的音と雑音の分

離結果を分析する。分析でテンプレートに対応する v_h が X_N に存在する可能性を計算できる。

最後に、すべての目的音の候補がそれぞれ X_N に存在すると仮定し、それに対応の X_N に存在する可能性を計算する。計算した結果の中で、妥当性が最も高い結果をとり、X_N に存在する目的音は以下の

$$v = arg\ max_v\{Eval\{Sep[X_N,\ C_v]\}\} \tag{2.1}$$

という数式から最後の認識結果 (Result of recognition) を得ることができる。

上記の「仮説・検証型」の音声認識法を実現するために、不可欠なツールは以下の 2 つである。

1. **認識用の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**
2. **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**

第3章　提案法の実装

本章では、提案法の実装方法を述べ、続いて本研究でツールとして用いられる Modified Restricted Temporal Decomposition (MRTD) や Non-negative Matrix Factorization (NMF) の概要を説明する。またこれらの手法の本研究での使い方および本研究の認識モデルの仕様を記述する。

3.1　提案法の実装方法

提案法の音声認識モデルを構築するため、必要なツールとして、本研究で用いられる方法は以下である。

- **認識用の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**

  本論文では、Nguyen らが提案した Modified Restricted Temporal Decomposition (MRTD) 合成法 [16] を合成器として用いる。MRTD 法は、音声のスペクトルパラメータの特徴を表す代表的なイベントベクトルを取り出し、これらのベクトルに対応する時間的な変化を表すイベントファンクションを計算する。この方法は、イベントベクトルとイベントファンクションの線形和で原音声を表す線形補間合成法である。イベントベクトルとイベントファンクションを調整することにより、表現された音声がコントロールできる。

- **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**

  本論文では、Lee らが提案した Non-negative Matrix Factorization (NMF) [17] 手法を目的音と雑音を分離する方法として用いる。NMF 手法は分離する音声を周波数領域の頻出パタンのベクトルを基底ベクトルと、各ベクトルの時間的なアクティブを表すアクティベーション行列に分解する。基底ベクトルと対応するアクティベーションの組み合わせをクラスタリングすることにより、音声分離ができる。本研究では、C_v の情報を基底ベクトルに与え、X_N の雑音と目的音を分離する。アクティベーションを解析することにより、目的音候補 v が X_N に存在する可能性が計算できる。さらに、NMF 手法は雑音と目的音のアクティベーションの組み合わせを考慮する必要がないので、処理高速化が期待できる [18]。

MRTD と NMF の紹介、および、本研究の認識モデルの実装については、次から述べる。

3.2 Modified Restricted Temporal Decomposition (MRTD)

3.2.1 MRTD の概要

MRTD [16] 手法は、線形補間の考えに基づき音声を表現する方法である。この方法は Temporal Decomposition (TD) [19] および Restricted Temporal Decomposition (RTD) [20] と同様、共同発音効果の線形モデルに基づく。式 3.1 に示したように、MRTD は原音声のスペクトルに関連性があるイベントターゲットと、時間的に重畳するイベントファンクションの線形結合で近似する方法である。MRTD 手法は RTD のイベントファンクションの計算法を改善し、RTD より高い精度で表現ができる [16]。

$$\hat{y}(n) = \sum_{k=1}^{K} \alpha_k \phi_k(n), \quad 1 \leqslant n \leqslant N \tag{3.1}$$

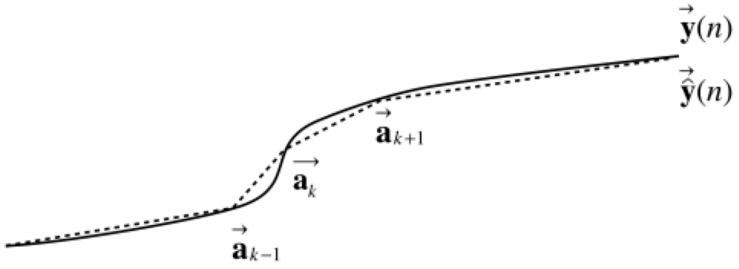
$\hat{y}(n)$ は近似する原音声のスペクトル関連量である。α_k と $\phi_k(n)$ は、k 個目のイベントターゲットとイベントファンクションである。K が増加するとともに、イベントターゲットとイベントファンクションの個数が増加し、$\hat{y}(n)$ の近似精度が高くなる。極端な場合には、$K = n$ になると、$\hat{y}(n)$ が完全に原音声と同等になる。

イベントターゲットとイベントファンクションのイメージを図 3.1 に示す。(a) に示すのはスペクトルパラメータ域で、原音声のスペクトルパラメータ $y(n)$ とその近似 $\hat{y}(n)$ である。α_{k-1}、α_k と α_{k+1} は 3 つの隣接のイベントターゲットである。このイベントターゲットは、$y(n)$ の中に特徴を持つベクトルである。(a) の中に、TD 手法は多次のスペクトル空間で 1 つ発話を複数のブレイクポイントで分析する。これらのブレークポイントがそれぞれのイベントに対応する。イベントターゲットに対応するイベントファンクションをかけ、その積を足し合わせると、$y(n)$ の近似 $\hat{y}(n)$ となる。(b) に示したのはイベントファンクションのイメージ図である。すべてのイベントファンクションの和が 1 となるという制約が TD 法にある。隣接のイベントファンクションの $\phi_k(n)$ と $\phi_{k+1}(n)$ だけは、時間軸上重畳である。

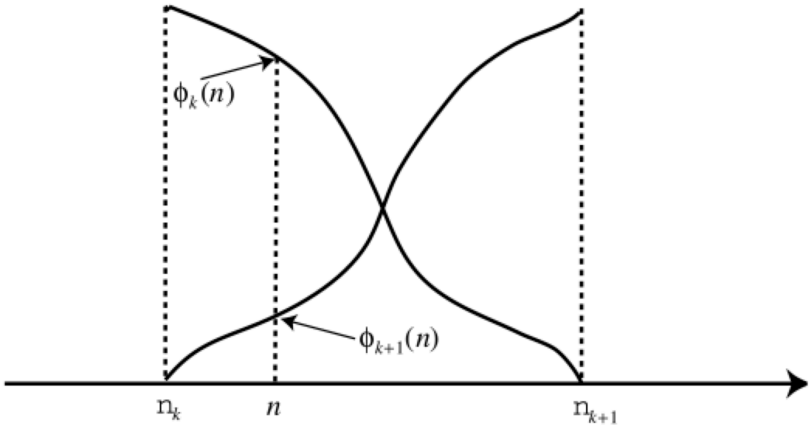
上記の制約によって、式 3.1 は以下のように書き換える：

$$\hat{y}(n) = \alpha_k \phi_k(n) + \alpha_{k+1}(1 - \phi_k(n)), \; n_k \; \leqslant n \leqslant n_{k+1}$$

MRTD で音声を表現するときには、まず原音声スペクトルパラメータ $y(n)$ の特徴を代表できるイベントターゲット α_k の位置を推定する。続いて、α_k の値を用いて対応するイベントファンクション $\phi_k(n)$ を計算する。図 3.2 で示したのは、MRTD を用いて計算された日本語発話の /shimekiri ha geNshu desuka/ のイベントファンクションである。図から分かるように、隣接のイベントファンクションだけが時間上で重畳し、イベントファンクションの和がいつも 1 となる。このように、プレークポイントの分析より、スペクトルパラメータをイベントに変換し、イベントターゲットとイベントファンクションで音声を表現できる。



(a)



(b)

図 3.1: イベントターゲットとイベントファンクションのイメージ

イベントターゲットの位置推定については、スペクトルパラメータの遷移性を表す spectral feature transition rate (SFTR) [21] を用いる。SFTR の極小値となる時間に対するベクトルが、音声の最も安定しているポイントである。これらのポイントでは、音声の特徴を持つ時間点である。このため、これらのポイントでのスペクトル関連量をイベントターゲットとして選択する。

SFTR の計算方法について例を挙げる。例えば、$y(n) = [y1(n)y2(n)...yI(n)]T$ がスペクトルパラメータの第 n 個のフレームである。I が一フ レームのパラメータ数であり、$y_i(n)$ が第 i 個のパラメータである。第 n 個のフレームに基づいて、$[n - M,\ n + M]$ サイズ の窓がかけている。以下の数式 3.2 と 3.3 をもちいて、SFTR を計算する。

$$SFTR: \quad s(n) = \sum_{i=1}^{P} c_i(n)^2, \quad 1 \leqslant n \leqslant N \tag{3.2}$$

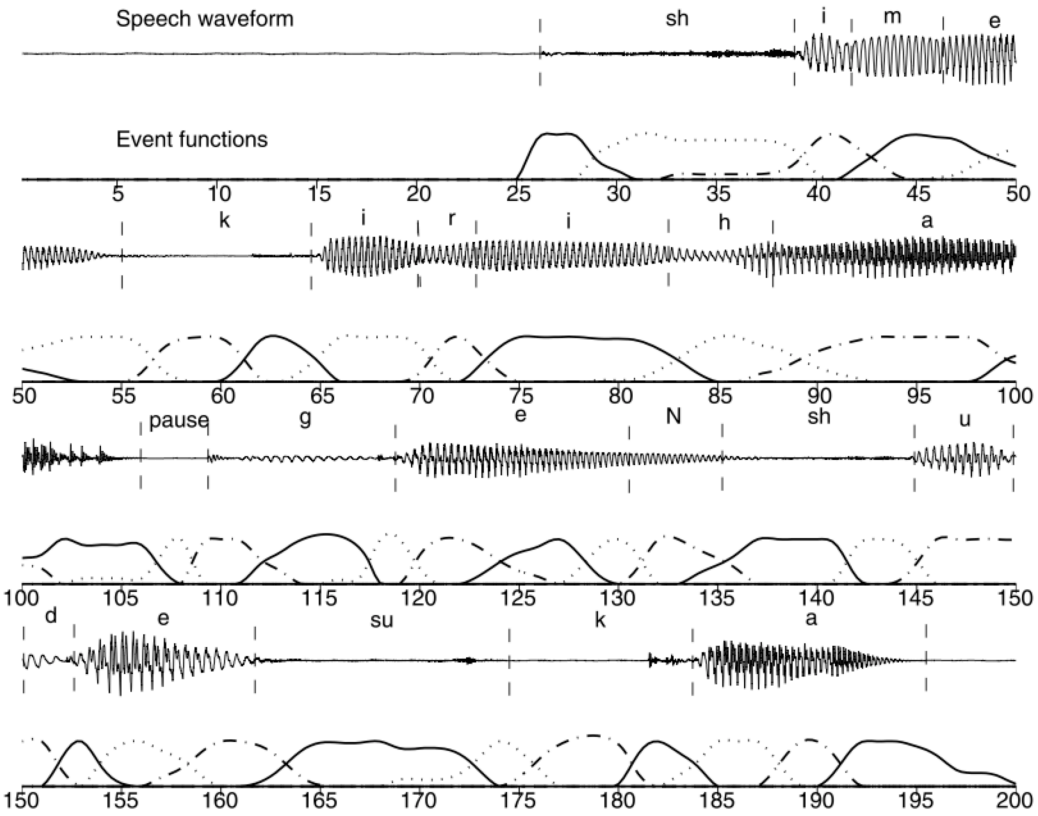


図 3.2: MRTD により計算されたイベントファンクションの例

$$c_i(n) = \frac{\sum_{m=-M}^{M} m y_i(n+m)}{\sum_{m=-M}^{M} m^2}, \quad 1 \leqslant i \leqslant P \tag{3.3}$$

式 3.2 と 3.3 で、P はスペクトルパラメータの次数であり、SFTR を計算する窓のサイズは $2M$ である。SFTR の極小値にあたるパラメータベクトルは、イベントターゲットとして初期化される。窓サイズ $2M$ が減少すると、SFTR が細かく計算され、極小値にあたるポイントが多くなる。逆に、極小値にあたるポイントが少なくなる。すなわち、M がイベントターゲットの数を影響する唯一の変数となる。

図 3.3 に示されたのは/a i kya ku/の音声の SFTR 図である。SFTR 図の上には、/a i kya ku/のスペクトルパラメータ（MFCC）である。図から分かるように、スペクトルパラメータの極小値にあたるポイントはイベントターゲットの位置として選択される。

イベントターゲット α_k が初期化されて後、それに対応するイベントファンクションの計算ができる。計算する方法は下記の式で表される。これらの公式で、音声の表現ができる。

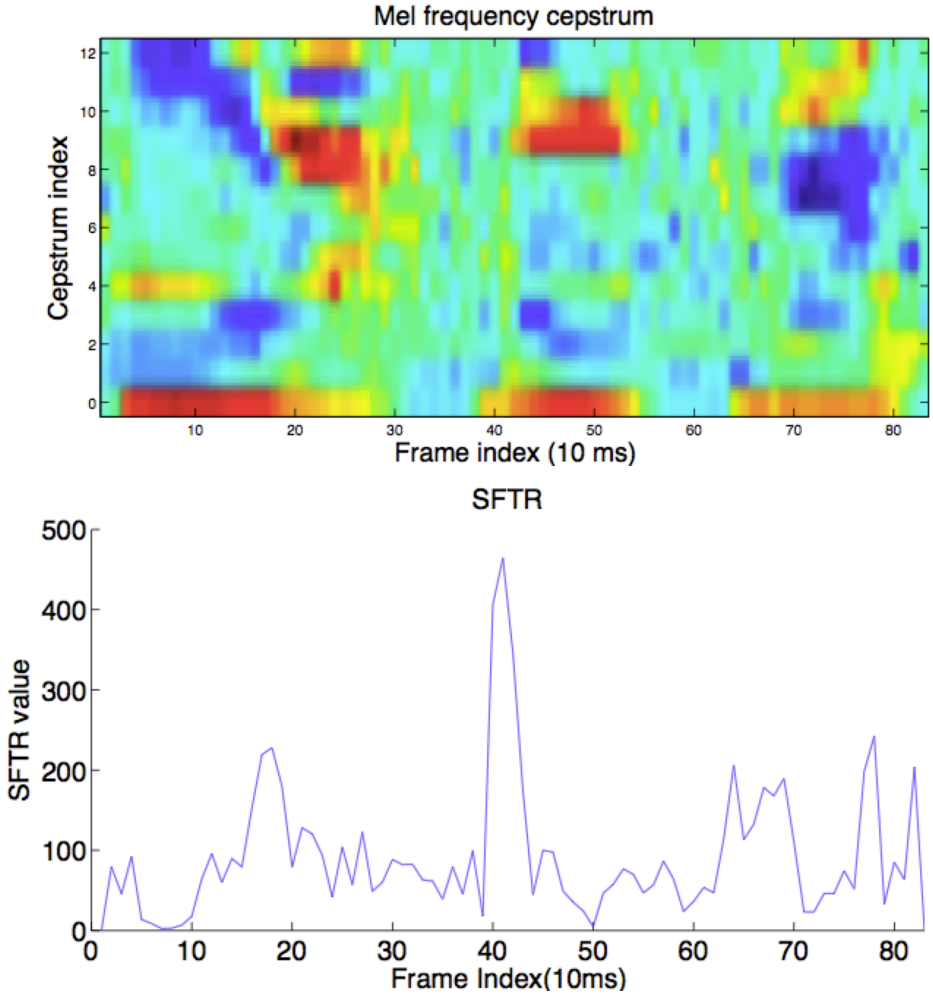


図 3.3: SFTR の一例

$$\phi_k(n) = \begin{cases} 1 - \phi_{k-1}(n), & if \;\; n_{k-1} < n < n_k \\ 1, & if \;\; n = n_k \\ min\left(\phi_k(n-1), max\left(0, \hat{\phi}_k(n)\right)\right), & if \;\; n_k < n < n_{k+1} \\ 0, & otherwise \end{cases} \tag{3.4}$$

$$\hat{\phi}_k(n) = \frac{\langle (y(n) - a_{k+1}), (a_k - a_{k+1}) \rangle}{\|a_k - a_{k+1}\|^2} \tag{3.5}$$

Nguyen らの提案した MRTD [16] では、線スペクトル対（LSF）をスペクトルパラメータとして扱った。しかし、本研究では雑音環境下での音声認識を行うため、LSF が雑音の影響を受けやすいので、音声認識用のパラメータとしては不適切である。本研究においては、雑音の影響と線形補間性を考慮する上で、メル周波数ケプストラム係数（MFCC）

をスペクトルパラメータとして扱う。本研究では (Hidden Markov Model Toolkit) HTK [22] の MFCC 生成法を参考とし、音声データの MFCC を計算する。

3.2.2 MRTD を用いた音声表現

メル周波数ケプストラム係数（MFCC）は下記のメリットがある：

1. ヒトの聴覚上重要な周波数成分が引き伸ばされて、ケプストラム全体における割合が増える。2. メルフィルタバンクを通すことで、メル周波数ケプストラムの特徴量の次元数が減り、計算の負荷が減る。

また、MFCC がよい線形補間性を持つため、本研究で音声データの MFCC を MRTD により表現し、テンプレートの生成に用いる。

MFCC の各パラメータの設定は表 3.1 で示されている。

表 3.1: MFCC に関する変数の設定

Parameter	Value	Significance
T_w	25	analysis frame duration (ms)
T_s	10	analysis frame shift (ms)
α	0.97	preemphasis coefficient
R	300 - 3700	frequency range to consider
M	20	number of filterbank channels
C	13	number of cepstral coefficients
L	22	cepstral sine lifter parameter

上記の条件に基づき、本研究では MFCC を計算する。SFTR を計算する際に、表現された音の歪みを軽減させるため、本研究では多数の実験を行い、M を 30 ms に設定した。原音声の MFCC を用いて、MRTD で音声を表現した。。図 3.4 に、MRTD で表現した音声の一例を示す。図 3.4 の上の図は、/a i kya ku/の原音声の MFCC であり、真ん中の図は対応するイベントファンクションの図である。下は、/a i kya ku/の表現された音声の MFCC である。図 3.4 により、原音声が表現された音で表すことができ、表現された音の MFCC をテンプレートとして用い、音声認識を行うことが可能である。

MRTD 合成法を用い、本研究の音声認識テンプレートを生成することにより：

- 合成器でテンプレートのコントロールができ、複数回発話の変化を吸収することができる
- MRTD で音声データをイベントターゲットとイベントファンクションで保存することが可能で、膨大なテンプレートデータを用いる際に、データ圧縮が期待できる

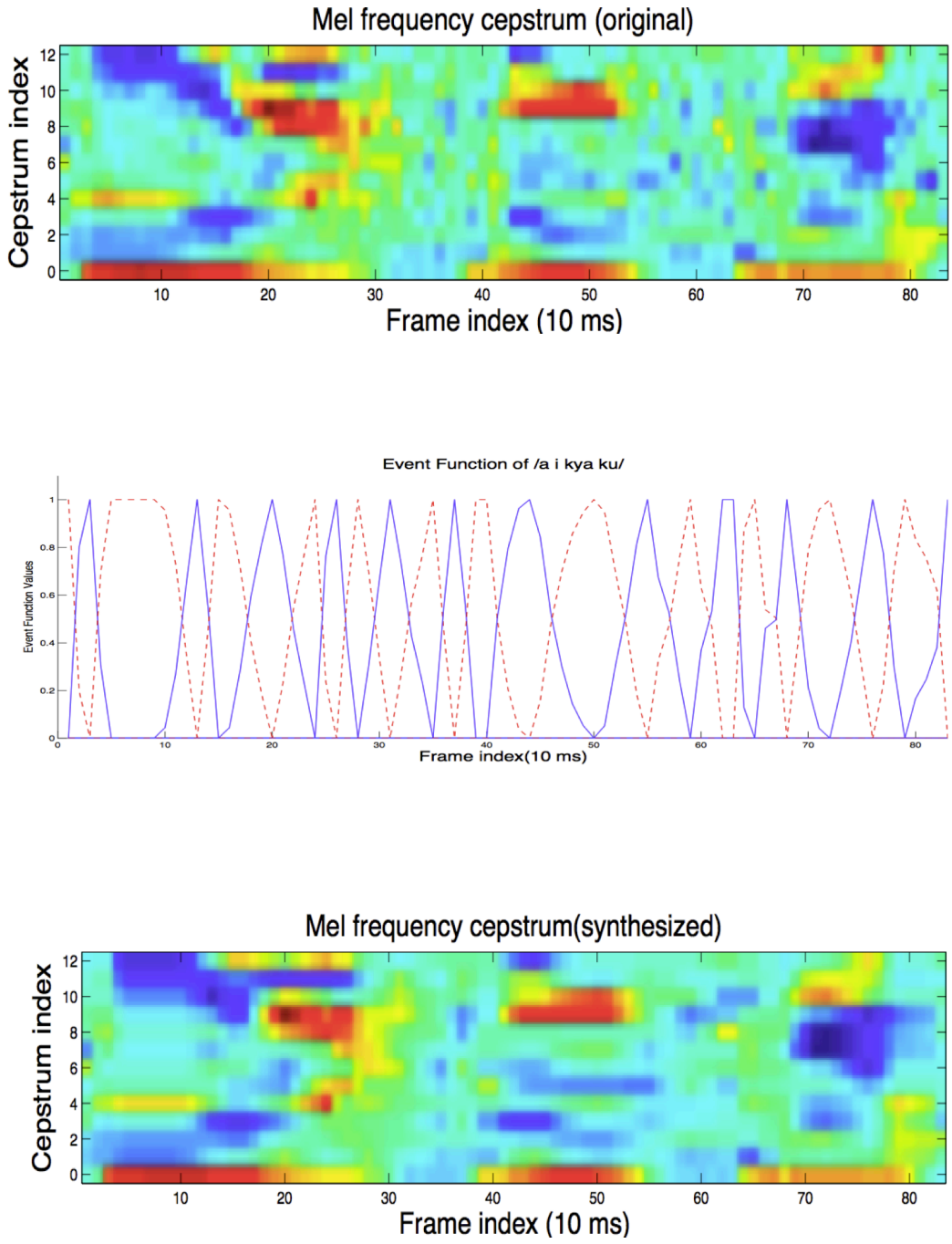


図 3.4: MRTD で表現された音声の一例

というメリットがある。

3.3 非負値行列因子分解 (NMF)

Lee らが提案した non-negative matrix factorization (NMF) [17] 手法は、様々な分野で注目を集めている。現実の世界では、パワースペクトル、画像値、頻度など、非負値で表されることが多い。また、主成分分析や独立成分分析で、構成成分を抽出することに役立つ場面が多い [23]。例えば、複数の音源の音響信号が混在する多重音のパワースペクトルから個々の音源のパワースペクトルを取り出すことができ、雑音除去や音源分離に役立てることができる。実際に、NMF は多重音に対する音源分離 [24] や背景雑音が存在する音声認識の研究 [25] にまで応用されている。

上記の理由で NMF 手法は、本研究で目的音と雑音をそれぞれのパワースペクトルの和の形で、分離する手法として用いられる。

3.3.1 NMF の概要

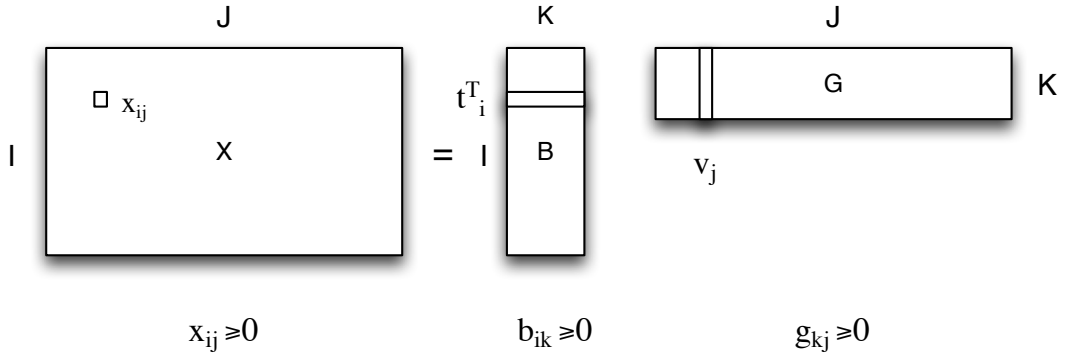


図 3.5: NMF のコンセプト

NMF とは、図 3.5 で示した 1 つの非負値行列を 2 つの非負値行列に分解する方法である。与えられた $I \times J$ の行列 X に NMF の入力し分解する。分解される 2 つの行列は $I \times K$ の行列 B と $K \times J$ の行列 G である。式 3.6 に示すように、X が B と G の積で近似される。B 行列は、X の構成成分を抽出した頻出パタンを表す基底行列であり、G 行列は B のベクトルのアクティブを表すアクティベーション行列である。K は NMF の基底数であり、一般には解析する人が事前に決めておく。基底数 K の増加とともに、B と G の規模が拡大するが、X の推定精度は高くなる。3 つの行列のすべての要素が、$x_{ij} \geqslant 0, b_{ik} \geqslant 0, g_{kj} \geqslant 0$ という非負制約がある。

$$X \approx B \times G \tag{3.6}$$

$b_i = [b_{1i}, \cdots, b_{Ki}]^T$、$g_j = [g_{1j}, \cdots, g_{Kj}]^T$ とすると、これらの内積

$$b_i^T g_j = \sum_{k=1}^{K} b_{ik} g_{kj} \tag{3.7}$$

は x_{ij} と等しくなるべき値となる。このため、入力行列 X のベクトル x_j は、式 3.8 で示したように基底行列の各ベクトルの重み付けの和となる。

$$x_j = \sum_{k=1}^{K} g_{kj} b_k \tag{3.8}$$

上記で述べたように、NMF 手法はある非負行列の頻出パターンの非負行列と、それに対応するアクティベーションの非負行列を抽出する。2 つの非負値行列を用いて、その積で入力した行列 X に近似する。

3.3.2 NMF の距離尺度

図 3.5 のように、X を近似分解するときに、一般には誤差が発生する。そのため、行列 X と近似結果の $B \times G$ の距離 $D(X,\ BG)$ を定義し、この距離を最小化すること必要がある。NMF で広く用いられる距離は, Eu: Euclid 距離の 2 乗、KL: 一般化 Kullback-Lerbler divergence、IS: Itakura-Saito divergence の 3 種類である [26]。それぞれ、

$$D_*(X,\ BG) = \sum_{i=1}^{I} \sum_{j=1}^{J} d_*(x_{ij},\ b_i^T v_j)$$

と定義されている。これらの距離 d_* は以下の形となる。

$$d_{Eu}(x_{ij},\ b_i g_j) = (x_{ij} - b_i^T g_j)^2$$

$$d_{KL}(x_{ij},\ b_i g_j) = x_{ij} log \frac{x_{if}}{b_i^T v = g_j} - x_{ij} + b_i^T g_j$$

$$d_{IS}(x_{ij},\ b_i g_j) = \frac{x_{ij}}{t_i^T v_j} - log \frac{x_{ij}}{t_i^T v_j} - 1 \tag{3.9}$$

これらの距離の中で、Euclid 距離は、x_{ij} と $b_i g_j$ の距離が 0 となる値を中心に対称である。一方で、KL 距離と IS 距離は非対称であり、値が大きくなりすぎることは許容されるが、足りないことには敏感である。また、IS 距離はスペクトルのピークを重視した距離であり、ホルマントの一致度をはかりやすい。このため、IS 距離は音声の処理に適切であり、本研究 NMF では IS 距離を用いることにした。

3.3.3 NMF の更新アルゴリズム

$D_*(X,\ BG)$ を最小化する NMF のアルゴリズムは、多数研究されている。本研究では、広く用いられている Multiplicative update rules で NMF アルゴリズムを実行する。与えられた行列の $(i,\ j)$ 成分 x_{ij} と等しくなるべき内積の値を $\hat{x}_{ij} = b_i^T g_j$ とする。IS 距離の更新式は下記の通りである：

$$b_{ik} \leftarrow b_{ik} \sqrt{\frac{\sum_j \frac{x_{ij}}{\hat{x}_{ij}} \frac{g_{kj}}{\hat{x}_{ij}}}{\sum_j \frac{g_{kj}}{\hat{x}_{ij}}}}$$

$$g_{kj} \leftarrow g_{kj} \sqrt{\frac{\sum_i \frac{x_{ij}}{\hat{x}_{ij}} \frac{b_{ik}}{\hat{x}_{ij}}}{\sum \frac{b_{ik}}{\hat{x}_{ij}}}} \tag{3.10}$$

ランダムな非負値で初期化した行列 B と G にこれらの更新式を何回か繰り返し適応することより、$D_{IS}(X,\ BG)$ が縮まっていく。このプロセスにより、分解後の行列 B と G が得られる。このように、更新式を用いて、更新する前の b_{ik}（あるいは v_{kj}）の値に別の値をかける更新形式は、Multiplicative update rule と呼ばれている。

この Multiplicative update rule を用いて、非負値行列 X の構成成分を抽出し、B と G で表現できる。目的音と雑音の構成成分が異なるため、本研究のコンセプトに基づき、目的音の特徴量を既知情報として扱い、NMF で目的音と雑音の構成成分を分離することが可能になる。

3.4 MRTD と NMF を用いた音声認識法

3.4.1 はじめに

NMF では非負制約があるため、スペクトルグラム（振幅スペクトルやパワースペクトル）が NMF 手法でパラメータとしてよく用いられる。本研究では、雑音と音声が独立であり、時間領域では振幅の加法性の性質を持っている。このため、フーリエ変換をすると雑音と目的音のスペクトルグラムも加法性の性質を持っている。NMF で入力スペクトルを分解し、基底ベクトルと対応するアクティベーション対は 1 つのコンポーネントとなる。このコンポーネントはある音源に属し、この音源に属するすべてのコンポーネントの重み付きの和で計算すれば、結果がこの音源のスペクトルとなる。このため、音源に属するコンポーネントのクラスタリング制約条件があれば、NMF を用いた音源分離が可能である。

制約条件や雑音と目的音の音源分離法の実装を後節で述べる。

3.4.2 音声分離法のコンセプト

この節では、本研究の音声分離法のコンセプトについて述べる。まずに、入力の雑音音声のパワースペクトルが、下記の式のように目的音 S と雑音 N のパワースペクトルの和で構成できる。

$$X = S + N$$

続いて、目的音と雑音を分離するために、目的音のパワースペクトル S に単位行列 I をかけ、雑音のパワースペクトル N を NMF のフォームで分解し、$B_N \times G_N$ の形になる。そして、雑音音声 X が下記の式になる。

$$X = S \times I + B_N \times G_N$$

この式を行列の形に書き換えると、下記の式となる。

$$X = [\ S \mid B_N\ ] \times [\ I \mid G_N\ ]^T$$

この式と NMF のアルゴリズムに関係をつけると、$[\ S \mid B_N\ ]$ は基底行列で、$[\ I \mid G_N\ ]^T$ はアクティベーション行列に見なすことができる。この式を利用し、本研究のアルゴリズムを実行する際に、テンプレートからの情報 $\hat{S}_h$ を用いる。$\hat{S}_h$ は目的音候補 v_h のテンプレートである。下記の式のように、$\hat{S}_h$ を用い、S を入れ替わり、固定する。また、ほかの 3 つの要素を NMF のアルゴリズムに従い更新する。

$$X = [\ \hat{S}_h \mid B_N\ ] \times [\ \hat{I} \mid G_N\ ]^T$$

このプロセスで、テンプレートの固定により、目的音と雑音の成分が分離される。また、S と $\hat{S}_h$ が近ければ近いほど、$\hat{I}$ が単位行列に近くなる。このことにより、$\hat{I}$ がどれほど単位行列に近いを評価すれば、目的音候補 v_h が入力に存在する可能性の計算ができる。

このようなコンセプトにより、目的音と雑音の分離ができ、また、分離結果を評価し、音声認識ができる。

3.4.3 音源分離法の実装

図 2.1 の中の Separation の詳細を、音声分離法の実装のイメージとして、図 3.6 に示す。X_N (Input) に存在する目的音 v が、目的音候補 v_h であると仮定 (Hypothesis v_h) する。本研究は、この仮説の妥当性を検証することにより、音声認識ができるというコンセプトを持っている。このコンセプトは第 1 章で述べた。このコンセプトのキーポイントは、音声分離する際に目的音候補 v_h のテンプレートを MRTD で表現されたデータ (Synthesized Data) から取り出し、分離する際にテンプレート C_v (Template C_v) を利用することができる点である。

本研究では、NMF を音声分離法として用いた。仮説 (Hypothesis v_h) より v_h に対応するテンプレートを抽出し、既知情報として、X_N を分離する際に NMF の基底ベクトルの

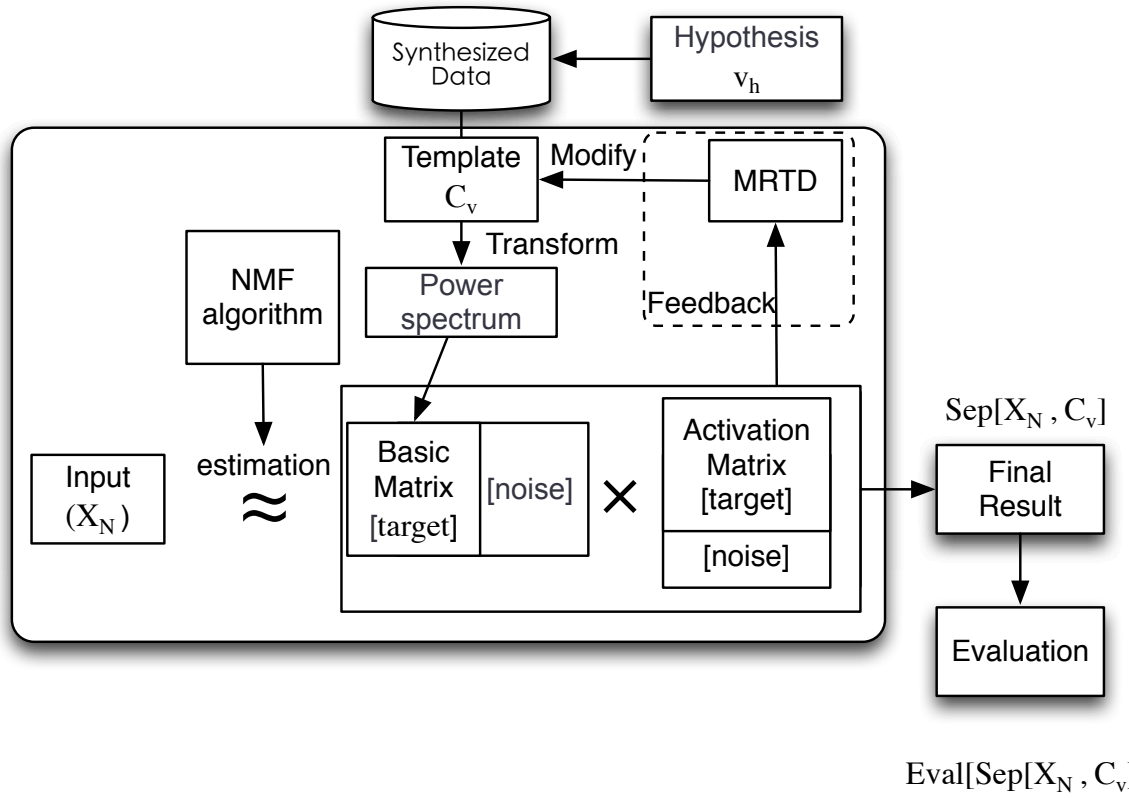


図 3.6: 音声分離法の実装

一部分 (Basic Matrix [target]) として固定した。NMF の非負制約があるため、テンプレートとして表現した MFCC をパワスペクトルに変換することが必要である。同じように、X_N は雑音音声であり、音声分離する前に X_N の MFCC を計算し、パワスペクトルへ変換する。変換したものを NMF アルゴリズム (NMF algorithm) の入力として、基底ベクトルの一部分 (Basic Matrix [target]) が固定された状況下で、音声分離を行う。前節で述べたように、NMF で音声分離をする際にクラスタリングの制約条件が必要である。基底ベクトルに固定された h_v のパワースペクトルが、その制約条件である。

具体的には、図 3.6 に示したように C_v のパワースペクトルを基底ベクトルの目的音部分 (Basic Matrix [target]) として固定した。式 3.10 より B のノイズ部分 (Basic Matirx [noise]) と G (Activation Matrix) を更新すれば、雑音と音声のパワースベクトルの加法性により、固定された部分に対応するアクティベーション行列 (Activation Matrix [target]) が、強制的に X_N に含まれる C_v のアクティブを表す。もし、目的音候補 v_h が X_N に存在すれば、C_v のアクティブを表すためには、アクティベーション行列の目的音部分 (Activation Matrix [target]) が近似単位行列となるべきである。雑音部分のコンポーネント (Basic Matrix [noise] と Activation Matrix [noise] の組み合わせ) は自由に更新されるため、X_N

に C_v と相似していない成分（雑音）が分離される。逆に、v_h が X_N に存在しなければ、C_v と相似していない成分（音声も含まれる）が雑音のコンポーネントに分離され、アクティベーションの目的音部分 (Activation Matrix [target]) は単位行列の形にならない。

このように、基底ベクトルを固定し NMF アルゴリズムによって、目的音と雑音の分離ができる。さらに、アクティベーション行列 (Activation Matrix) を解析すれば、X_N に存在する可能性の最も高い目的音候補が取り出せる。よって、この方法で、雑音環境下での音声認識ができる。

図 3.7 と 3.8 で示したのは、クリーンと 10 dB のピンクノイズ環境下でのアクティベーション例である。この例の中には、入力 X_N に含まれる目的音 v は日本語単語 /i ki o i/ である。 図 3.7 と 3.8 に、上のほうは目的音候補 v_h が /i ki o i/ 、すなわち目的音と仮定した目的音候補が一致とした状況である。下のほうは v_h が /jyu N ba N/ 、すなわち目的音と仮定した目的音候補が一致としていない状況である。これにより、クリーンな環境にもかかわらず、v_h が X_N に存在する状況にあたり、アクティベーション行列の目的音に対応する部分 (Activation Matrix [target]) が予測通りに対角行列に近づくことになった。一方、v_h が X_N に存在しない状況にあたり、アクティベーション行列の目的音に対応する部分 (Activation Matrix [target]) に、重みの分布が少なく、対角行列となっていない。図から分かるように、この方法を用い雑音の影響を受けずに、目的音候補 v_h が X_N に存在する可能性を判断できる。また、分離法としては X_N のパワースペクトルを C_v に相似している成分と相似していない成分に分離することができる。

今回の研究では、まだ次の項目が実現できていない。アクティベーション行列の目的音部分 (Activation Matrix [target]) の情報から、MRTD へフィードバックを与え、MRTD がフィードバックにより、テンプレート (Template C_v) を微調整する。そこで、目的音と雑音の分離結果をさらに単位行列へ近づけることを行う。今後、拡張を行う必要がある。

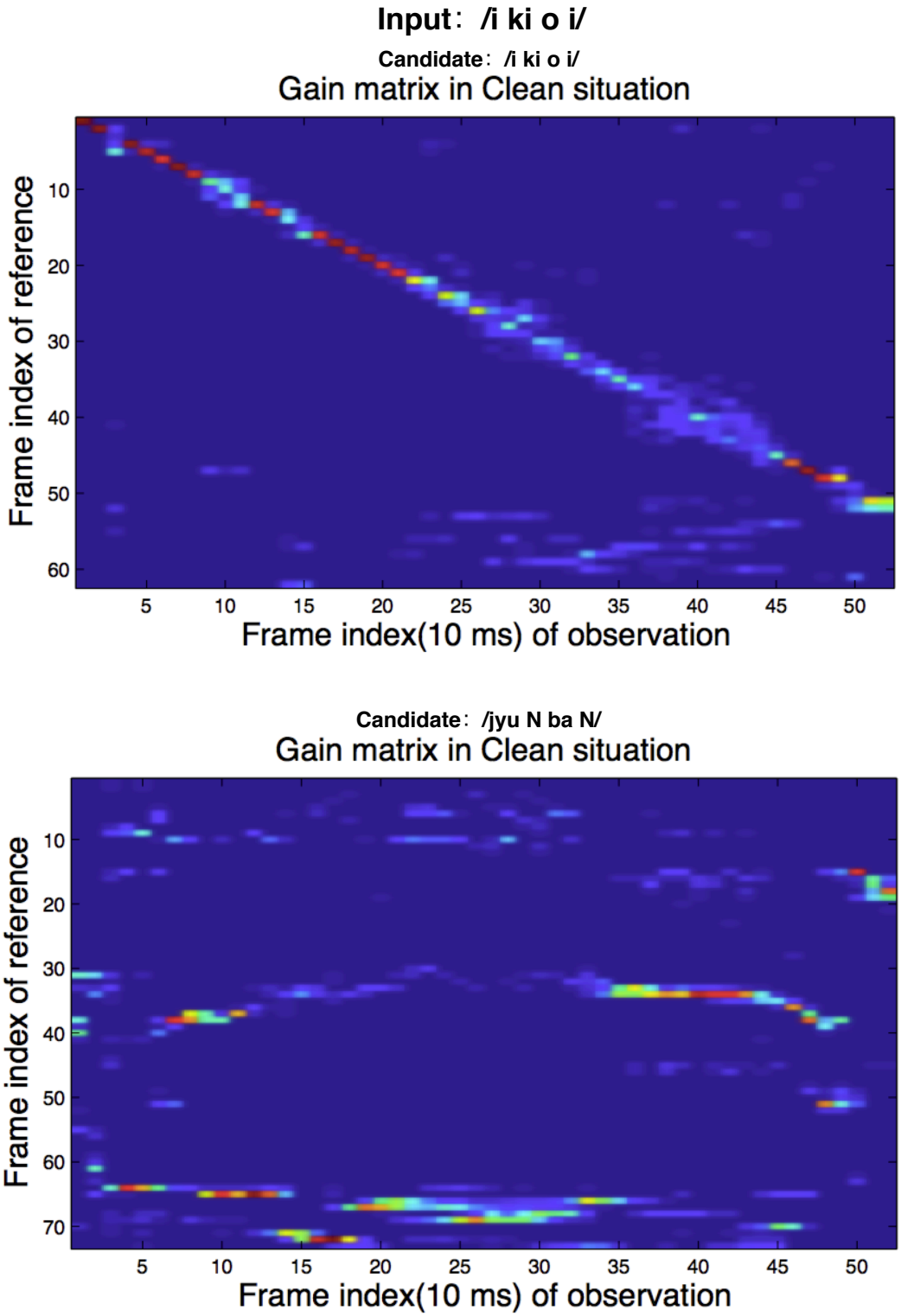


図 3.7: クリーンな環境下　入力：/i ki o i/ 候補：/i ki o i/（上)、 /jyu N ba N/（下）

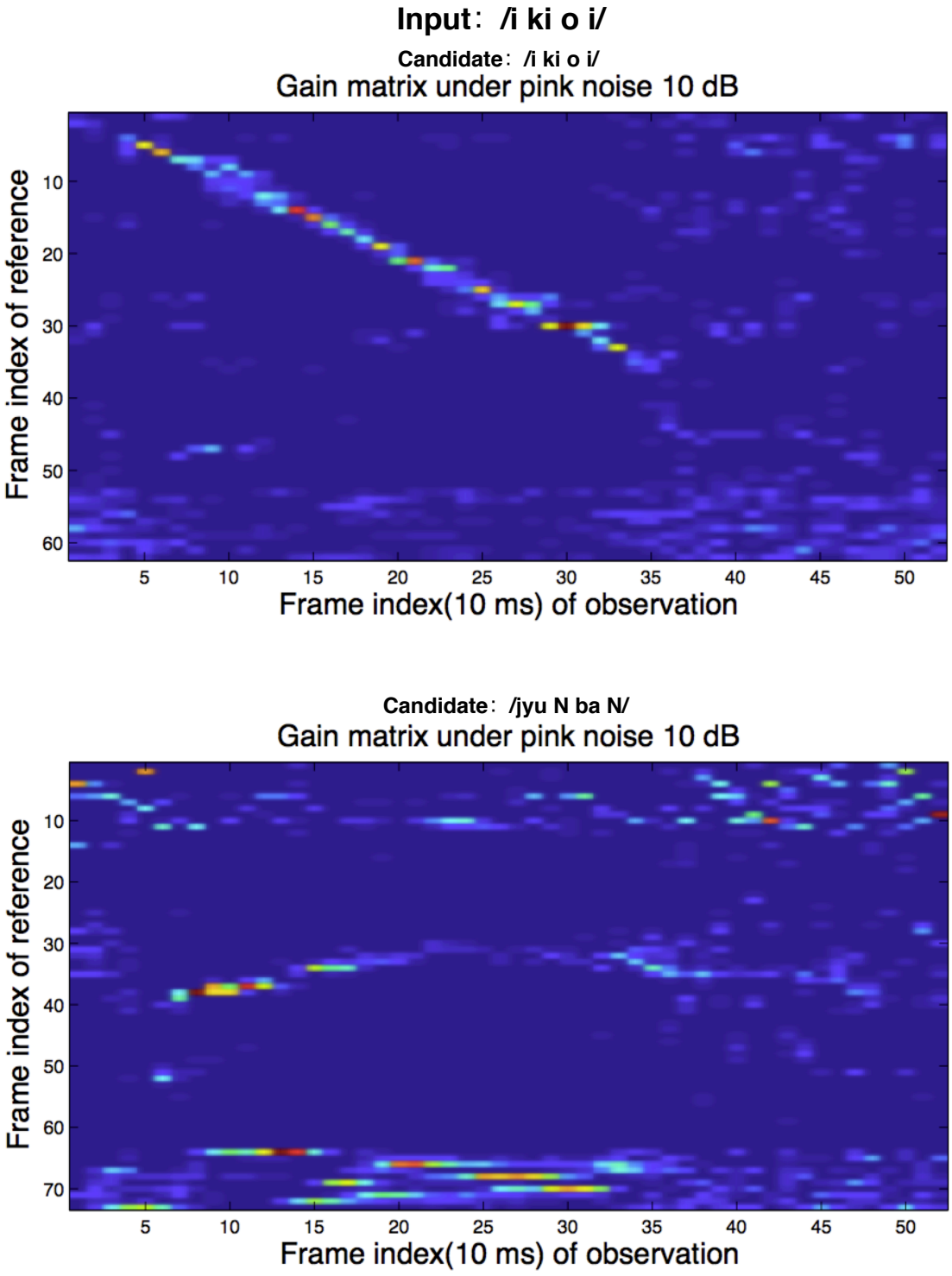


図 3.8: 雑音環境下　入力：/i ki o i/ 候補：/i ki o i/（上)、 /jyu N ba N/（下）

3.4.4 認識法の実装

図 2.1 に示したように、目的音を認識をするため、すべての目的音候補に対応する音声分離の結果を評価することが必要である。この評価の結果により、X_N に存在する可能性の最も高い目的音候補 v を認識する。

前節で述べたように、v_h が X_N に存在すると、アクティベーションの目的音部分が対角行列に近づく。さらに、v_h が X_N に相似すればするほど、この部分は単位行列に近づく。このため、アクティベーション行列の目的音部分の値の対角線付近の分布率を評価標準として用いた。図 3.9 に示したように、黄色の範囲内で多くの値が分布すると、この行列が単位行列に近くなる。すなわち、v_h が X_N に存在する可能性が高くなる。

この標準を用いて、本研究では NMF を用い、式 2.1 のコンセプトに基づく音声認識モデルが実装できた。本研究の方法を用いた音声認識性能の評価は、次の章で述べる。

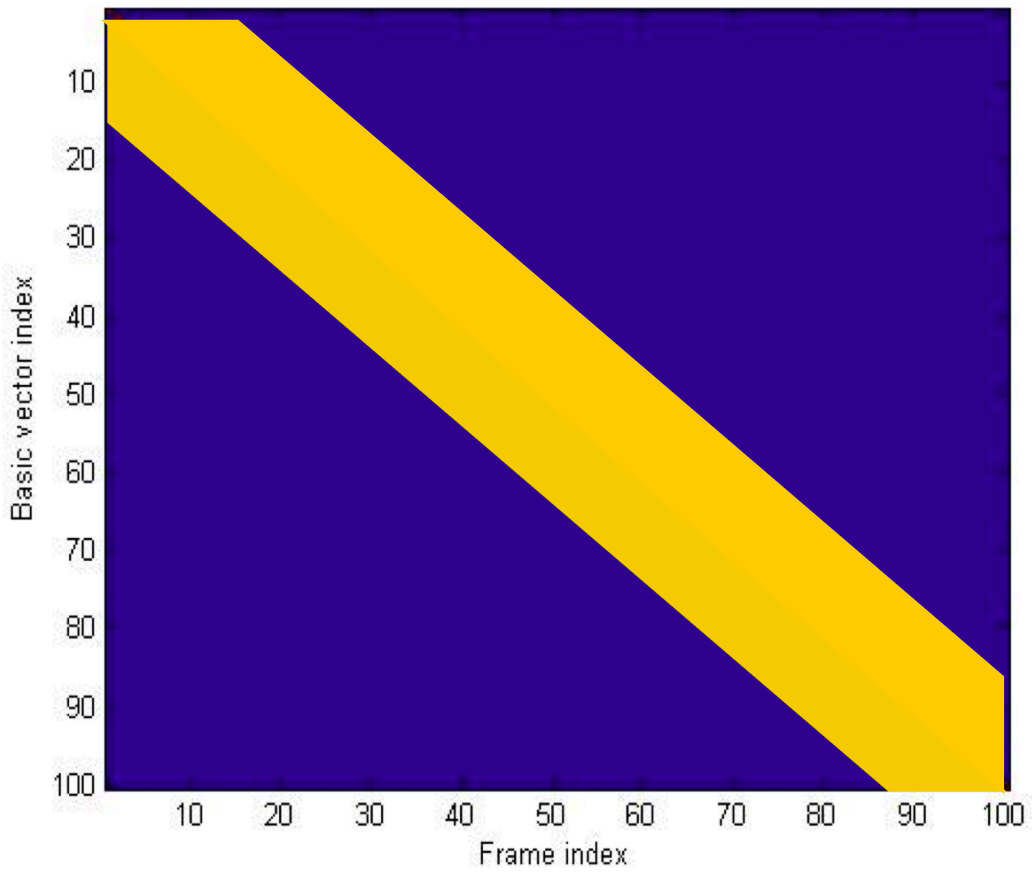


図 3.9: 分離結果の評価法

第4章　評価実験

本研究では、提案法の雑音環境下での有効性を評価する実験を行った。4 種類の雑音に対して、提案法における音声認識法と Dynamic Time Warp (DTW) [28] における音声認識法それぞれで単独語彙の音声認識を行った。続いて、それぞれの認識結果を比較した。本章では、その結果を報告する。

4.1　評価実験の目的

予備実験では、雑音環境下での音声認識精度を向上させるため、ASA のコンセプトに従い、MRTD と NMF を用いて、音声認識法を実装した。予備実験の結果により、この方法は目的音と雑音の分離により仮説・検証のプロセスで、入力雑音音声の X_N にどの目的音候補 v が存在するのかが判断できた。

本実験では、本研究の音声認識法が雑音への頑健性を持っていることを証明することである。本研究における仮説・検証型の音声認識法を用いて、典型的なテンプレート音声認識法 DTW と比較した。今回の認識手法は、まだ時間伸縮などのアルゴリズムが完成していないため、DTW の Dynamic Programing (DP) [29] アルゴリズムを用い、提案法と同じ条件下で入力目的音を認識する。その結果を提案法における認識結果と比較する。比較用の音声認識法の実装は、Ellis らのコード [31] を参照した。

4.2　評価実験の条件

本研究では MRTD と NMF を用いて実装した音声認識法の有効性を検証する。本研究では、実環境で使用可能の音声認識の第一歩として簡単な状況を仮定した。まず認識用のテンプレートを作成する原音声と入力音声が同一話者、同一発話と仮定する。異なる発話のスペクトルパラメータに多少の変化がある。しかし、本研究ではそれを無視し、コンセプトの有効性だけに注目した。今後の拡張では、フィードバックの情報により、MRTD でテンプレートを修正する方法を加え、異なる発話の変化を吸収する方法が可能と考えている。

本研究の予備実験で、ATR データベース A [30] の音韻バランス語の中に、4 モーラ 10 単語を音声認識データとして選び、4 種類の雑音環境下で音声認識を行った。その音声データが音韻バランス語であるため、各単語のばらつきが大きい。本研究のコンセプトに従う

ことにより、雑音と目的音を分離し、評価することで目的音候補を認識することが容易である。このため、本研究の音声認識法を用いて、よい認識率が得られた。

本実験を行うため、音韻バランス語より各単語のばらつきが小さい親密度了解度実験用データベース（F0W3）[31] を選択した。そのデータベースの中に、話者 “fto”が発話した 4 モーラ 100 単語を音声認識データとして選んだ。

雑音データに関して、本研究では 4 種類の雑音: white noise, pink noise, babble noise, factory noise を用いた。雑音は目的音の関係が加法関係であると仮定した。雑音環境は 0 dB, 10 dB, 20 dB とクリーンな環境と設定した。本研究においては、クリーンな環境、および 12 種類の雑音環境、計 13 種類の環境で認識実験を行った。それぞれの環境下で、100 単語をそれぞれ 1 回入力する。毎回の入力音声に対し、雑音環境と対応する SNR を計算し、ランダムに生成した雑音と音声を足し、その MFCC を計算した。最後に、得られた MFCC を本研究の認識法の入力とした。

本研究の比較実験の全体的なデータフロー図は図 4.1 である。Comparition の上には本研究の音声認識法であり、下には DTW における音声認識法である。本研究では MRTD を用いて、テンプレート (Template C_v) を表現した。また、MFCC を認識パラメータとして用いた。また、MRTD の有効性を検証するため、原音声の MFCC と表現した MFCC をテンプレートとして用い、予備実験を行った。その結果から、MRTD で表現したテンプレートを用いて、認識プロセスに影響が弱く、十分の有効性をしめした。本実験に、入力音声 X_N に存在しうる目的音候補 (Candidates) を決定し、テンプレート (C_v) を選んだ。NMF の非負性制約をあわせるために、テンプレートの MFCC からパワースペクトルに変換した。入力音声の MFCC も同じくパワースペクトルに変換した。

そして、本研究の音声認識法で目的音を認識するときに仮説を用い、目的音と雑音のパワースペクトルを NMF で分離し、分離の結果を評価 (Evaluation) する。すなわち、アクティベーション行列の対角線付近の重みの分布率が最も高い目的音候補を音声認識の認識結果 (Result of recognition) とする。比較 (Comparition) するため、DTW における音声認識法を実施した。この方法では、提案法と同じように入力 (Input) やテンプレート (Templates) が MFCC からパワースペクトルに変換した。テンプレートと入力の尤度計算より (Recognition DTW)、尤度の最も高い単語を認識した。

最後に、それぞれの認識結果とその比較は次の節で述べる。

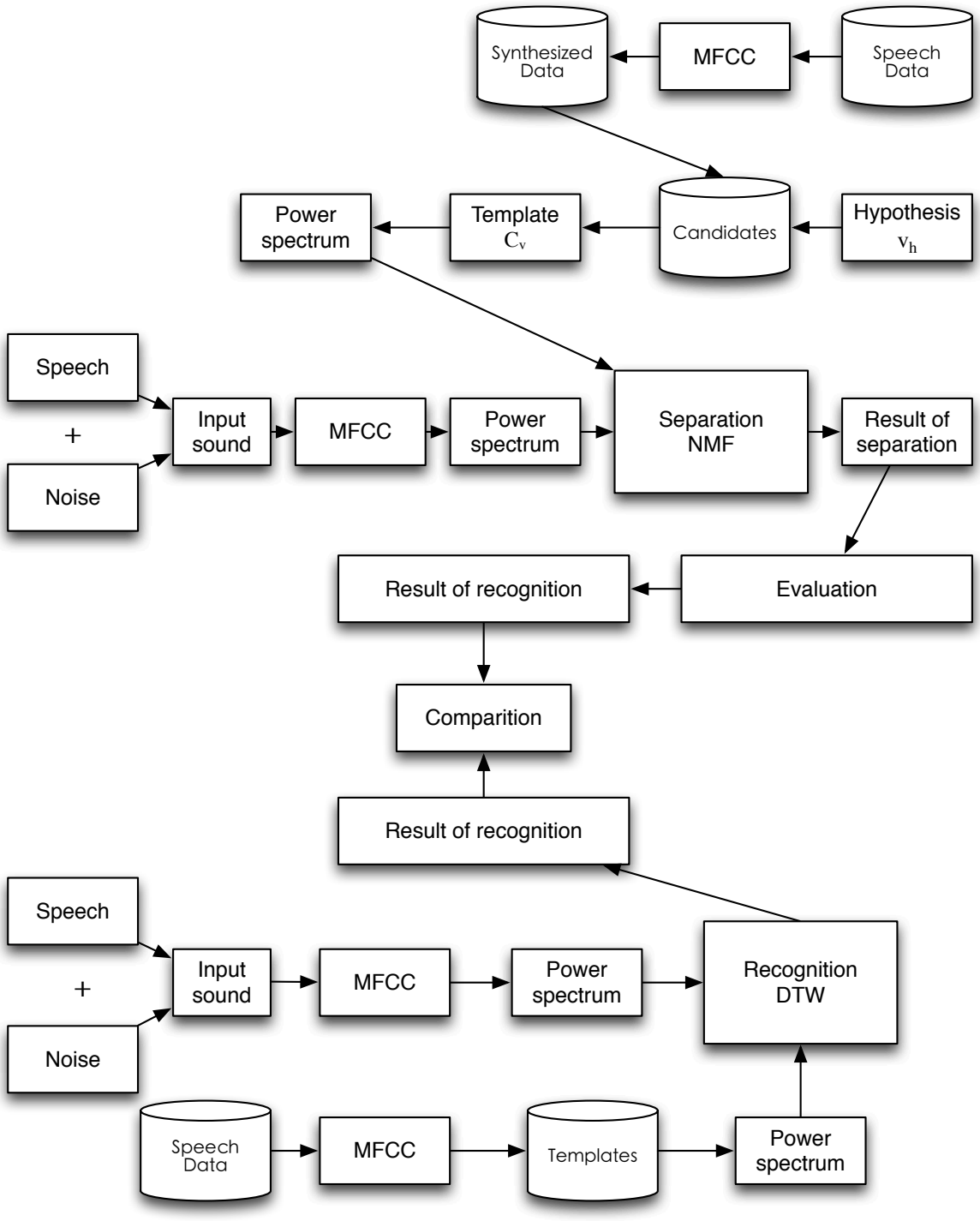


図 4.1: 比較実験のデータフロー

4.3 評価実験の結果の考察

提案法における音声認識法の結果と DTW における音声認識法の結果をそれぞれ図 4.2 と 4.3 に示した。認識率は：

$$\text{認識率} = \frac{\text{正しく認識された単語数}}{\text{入力単語数}} \times 100\%$$

と定義した。

DTW における音声認識法の結果で、雑音環境下で音声認識率が大幅に下がった。特に 0 dB の環境下で、わずか 20%の認識率しかない。しかし、本提案法における音声認識法の結果では、前処理や雑音モデルで雑音環境に適応せずに、0 dB の雑音環境に対しても約 80%の認識率が得られた。

また、処理時間については、本研究では ASA のコンセプトを用い 100 個の目的音候補の中から、1 つの最適な音声を認識する所用時間は、約 20-30 分である。一方、羽二生ら提案法を用いた音声認識法が 10 個の目的音候補の中から、1 つの最適な音声を選択する所用時間が 1 日以上である。

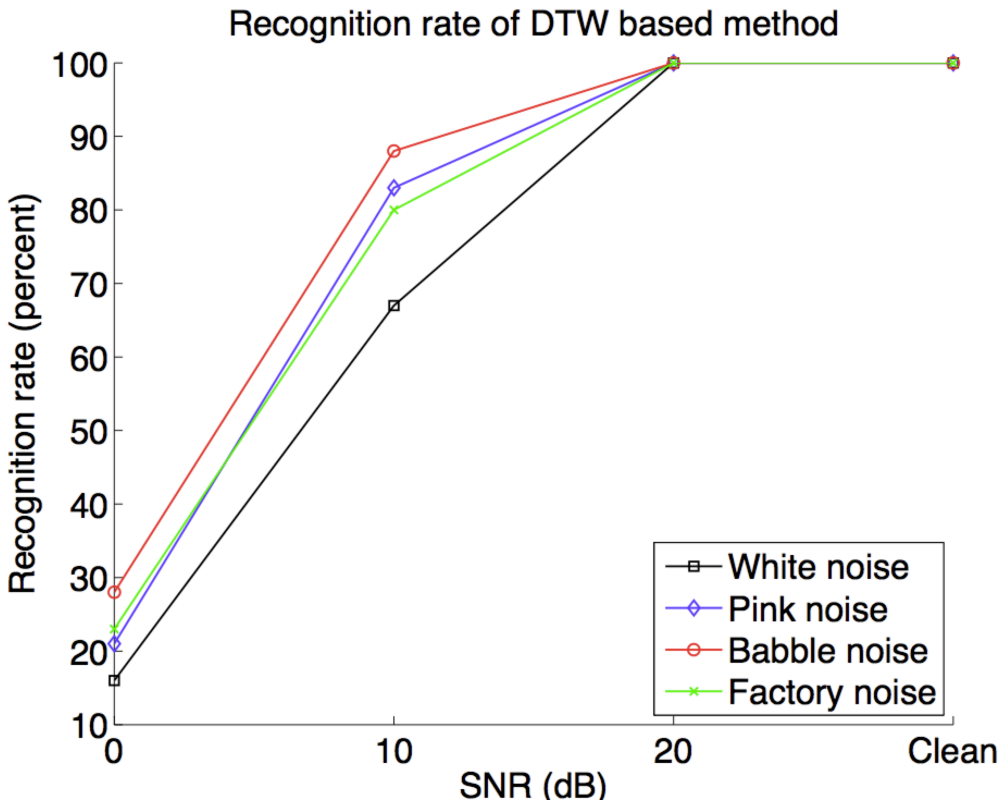


図 4.2: DTW における音声認識法の結果

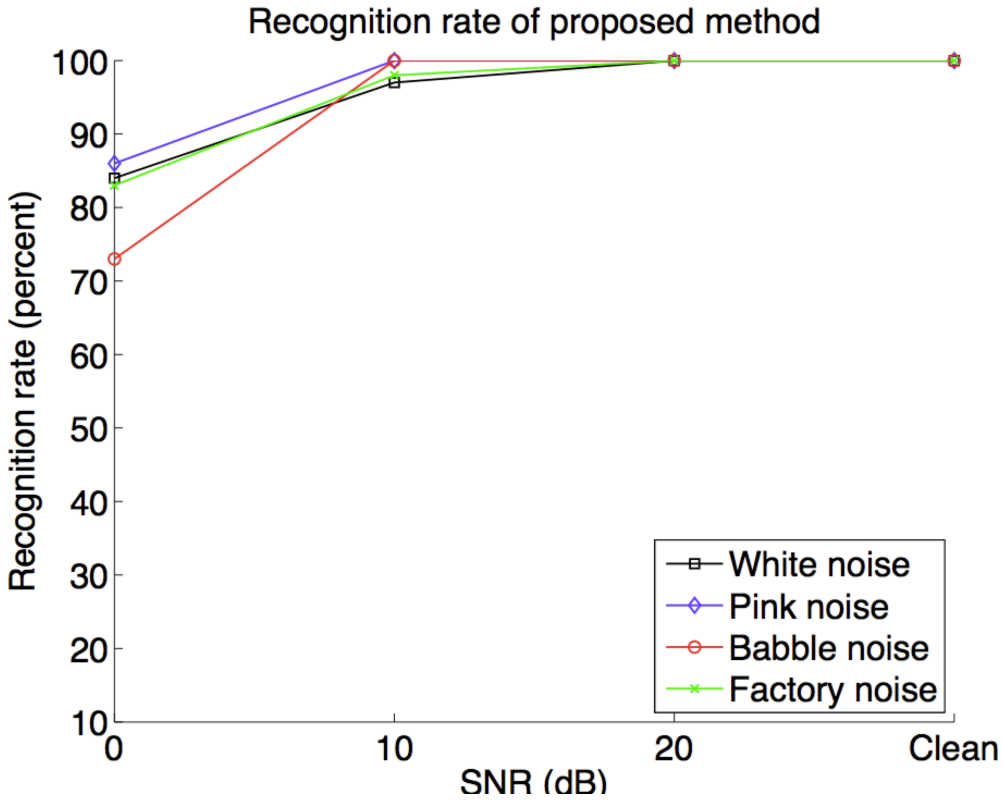


図 4.3: 提案法における音声認識法の結果

4.4 まとめ

本研究の提案法は、MRTD と NMF における音声認識手法である。この手法は ASA のコンセプトに従い、入力する雑音音声の中に含まれる目的音と雑音を分離することにより、単独語彙を認識することが可能である。

評価実験の結果から、本提案法における音声認識方法は、4 種類の雑音に対して頑健性を示した。さらに、羽二生らの手法に比べて計算上の利点を示した。

第5章　結論

本章では、本論文の内容を要約し、本研究から明らかになったことをまとめる。また、今後の課題について述べる。

5.1　まとめ

本研究では、音声認識システムが雑音の影響を受けやすい問題を解決するため、人間の聴覚能力を考慮し、ASA のコンセプトを用いて、音声認識の手法を提案した。このコンセプトにより、雑音と音声分離の結果により、最終的な分離音が目的音として妥当であるかどうかを判断した。これは、本提案法の音声認識の基準である。

上記のコンセプトを実現するため、解決すべき問題が以下の 2 つであった。

1. 認識用の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

この問題に対して、本論文では、MRTD 手法を合成器として用い、認識用のテンプレートを合成した。合成音がテンプレートとして、使えることを予備実験で証明した。また、この合成器を用いて、複数回発話の変化を吸収する拡張の可能性を示した。

2. 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

目的音候補と雑音を分離することができ、かつ処理の高速化が期待できるため、目的音と雑音を分離する方法として、本研究では NMF を用いた。予備実験の結果から、既知のテンプレート情報により、基底ベクトルの一部分を固定し、NMF のアルゴリズムを実行する場合で、目的音と雑音の分離ができる。また、アクティベーション行列の結果により、目的音候補が X_N に含まれる妥当性が検証できる。これにより、ASA のコンセプトに基づく音声認識手法の構築ができた。さらに、処理時間においては、同じく ASA のコンセプトに基づく例と比較すると、本研究の処理時間は圧倒的に短いことが分かった。

本研究では、これらの 2 つの問題を解決し、雑音環境下での音声認識手法の構築ができた。この認識手法の有効性を示すため、ホワイトノイズ、ピンクノイズ、バブルノイズと工場ノイズと FW03 の 4 モーラ単語を混合した音声から、目的音を認識するシミュレーションを行った。結果により、これらの雑音と日本語単語が加算された状況において本手法が有効であることを確認した。

上記の結果から、本研究は雑音に頑健である音声認識手法の第一歩として、その有効性を示した。

5.2 今後の課題

今後、本研究の提案法で単語認識することにより、雑音が混雑する車内や、駅などの環境下でキーワード認識技術を用い、コンピュータの音声操作への応用などが考えられる。また、MRTD で音素を表現し、本研究の提案法が音素認識まで拡張することが可能であるため、連続語彙の認識の可能性がある。さらに、単一話者の複数回の発話や不特定話者などに対応する可能性を持っている。以下に、これらの音素認識や連続語彙認識および多話者複数回発話に対応するために、必要な課題を示す。

音素認識の拡張

本研究はテンプレートを用いる単語を認識する方法である。音声認識では、テンプレートが単語であると、連続語彙などを認識することが困難である。今後の拡張として、MRTD で音素を表現しテンプレートを音素に拡張すれば、連続語彙などの認識が可能となる。この拡張を実現するため、MRTD の改良以外に入力音声を正しく音素に分割する方法が必要である。また、認識する際に、音素認識に対応する NMF 法の改良も必要となる。

連続語彙の拡張

音素認識ができることは連続語彙を認識するための必要な条件の 1 つである。また、連続語彙を認識するため、文法などの言語知識、すなわち言語モデルの応用が必要である。図 2.1 に示すように、言語モデルの知識が、目的音候補 v_h の系列を生成する時、役に立つ。ここで言語モデルの知識を用いて、正しい目的音候補の系列が生成でき、認識率の向上また処理の高速化が期待できる。

多話者複数回発話の適応

本研究のコンセプトにより、多話者や複数回発話の変化を吸収する方法は、膨大なテンプレートを用意することではなく、分離の結果から合成器を用いて、テンプレートを微修

正する方法である。このため、本研究では、目的音と雑音を分離する結果のフィードバックから、アクティベーション行列を単位行列に近づける修正法が必要となる。

ここで示した課題を克服することにより、本研究の提案手法は音声認識が利用できる状況を現在より大きく拡張する可能性を持っている。

謝辞

本研究を進めるにあたり、多大なる御指導ならびに御鞭撻を賜りました赤木 正人 教授に深く感謝致します。本研究を進めるにあたり、日頃から熱心な御指導ならびに御鞭撻を賜りました鵜木 祐史 准教授に心より感謝致します。本論文を作成するにあたり、貴重な時間を頂、熱心な指導を賜りました寺朱美先生に心より感謝致します。

そして、日頃から数多くの議論と激励をいただいた赤木研究室の諸先輩方に厚く御礼を申し上げます、また、本研究の遂行にあたり多面にわたり御協力いただいた音情報処理学講座の皆様に感謝致します。

最後に、本学での研究生活を支え、温かく見守ってくれた両親に心から感謝致します。

参考文献

[1] Wikipedia, “音声認識”, http://ja.wikipedia.org/wiki/音声認識.

[2] 鈴木陽一, 赤木正人, 伊藤彰則, 佐藤洋, 苣木禎史, 中村健太郎, “音響学入門”, pp.88-92, 2011.

[3] J. Benesty, M. M. Sondhi, Y. Huang(Eds), “Springer Handbook of Speech Processing”, pp.521-535, 2007.

[4] J. Benesty, M. M. Sondhi, Y. Huang(Eds), “Springer Handbook of Speech Processing”, pp.653-664, 2007.

[5] B. A. Dautrich, L. R. Rabiner, and T. B. Martin, “The effects of selected signal processing techniques on the performance of a filter-bank based isolated word recognizer”, Bell Systems Technical Journal 62,1311, 1983.

[6] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. Acoustics, Speech Signal Process. ASSP-27, pp.113-120, 1979.

[7] R. Zelinski, “A microphone array with adaptive post-filtering for noise reduction in reverberant rooms”, IEEE, Acoustics, Speech, and Signal Processing, ICASSP-88., pp. 2578-2581, 1988.

[8] M. J. F. Gales and S. J. Young, “Robust speech recognition using parallel model combination”, IEEE trans. on Speech and Audio Processing, Vol.4, pp.352-359, 1996.

[9] F. Martin, et al., “Recognition of noisy speech by composition of hidden Markov models”, Proc. Eurospeech’ 93, pp.1031-1034.

[10] 赤木正人, 羽二生篤, “音声の知覚と認識：人は脳で音声を聞く．機械は？”, 日本音響学会論文集, 2011, pp.1725-1728.

[11] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears”, J. Acoust. Soc. AM., pp. 975-979, 1953.

[12] A. S. Bregman, “Auditory scene analysis: The perceptual organization of sound”, MIT Press, 1990.

[13] A. Haniu, M. Unoki and M. Akagi, “A study on a speech recognition method based on the selective sound segregation in noisy environment”, NCSP2005, 403-406, 2005.

[14] 羽二生篤，鵜木祐史，赤木正人, “ヒトの聴覚情報処理過程を考慮した音声認識モデル”，電子情報通信学会技術報告，SP2009-33, 2009.

[15] M. Unoki, M. Akagi, “A method of signal ex- traction from noisy signal based on auditory scene analysis”, Speech Communication 27, pp.261-279, 1999.

[16] P. C. Nguyen, T. Ochi and M. Akagi, “Modified Re- stricted Temporal Decomposition and Its Application to Low Rate Speech Coding”, IEICE TRANSACTIONS on Information and Systems, E86-D(3):397-405, 2003.

[17] D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization”, Adv. Neural Inf. Process. Syst., pp.556-562, 2000.

[18] S. J. Rennie, J. R. Hershey and P. A. Olsen, “Single- Channel Multitalker Speech Recognition”, IEEE Signal Processing Magazine, pp. 66-80, 2010.

[19] B. S. Atal, “Efficient coding of LPC parameters by temporal decomposition”, Proc. ICASSP’83, pp.81-84, 1983.

[20] S. J. Kim and Y. H. Oh, “Efficient quantization method for LSF parameters based on restricted temporal decomposition”, Electron. Lett., vol.35, no.12, pp.962-964,1999.

[21] A. C. R. Nandasena and M. Akagi, “Spectral stability based event localizing temporal decomposition”, Proc. ICASSP’98, pp.957-960, 1998.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, “The HTK Book”, Version 3.4, pp.73-90, 2009.

[23] 亀岡弘和, “非負値行列因子分解の音響信号処理への応用”, 日本音響学会誌, vol.68, no.11, pp.559-565, 2012.

[24] Tuomas Virtanen, “Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria”, IEEE Transactions on Audio, Speech, and Language Processing, vol.15, no.3, 2007.

[25] S. Nakano, K. Yamanoto, and S. Nakagawa, “Fast NMF based approach and improved VQ based approach for speech recognition from mixed sound”, Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), pp.1-4, 2012.

[26] 澤田 宏, “非負値行列因子分解 NMF の基礎とデータ／信号解析への応用”, 電子情報通信学会誌 vol.95, no.9, pp.829-833, 2012.

[27] L. Muda, M. Begam and I. Elamvazuthi, “Voice Recognition Algorithms using Mel-Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, Journal of computing, vol.2, issue 3, pp.136-143, 2010.

[28] D. Ellis, “Dynamic Time Warp (DTW) in Matlab”, http://www.ee.columbia.edu/ln/labrosa/matlab/dtw/, 2003.

[29] H. Sakoe, Nippon Electric Company, Limited, Kawasaki, Japan, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, Acoustics, Speech and Signal Processing, IEEE Transactions. vol.26, Issue.1, 1978.

[30] 武田一哉, 匂坂芳典, 片桐滋, 阿部匡, “研究用日本語音声データベース”, エイ・ティ・アール自動翻訳電話研究所.

[31] A. Shigeaki, K. Kondo, S. Sakamoto, Y. Suzuki, “Speech Data Set for Word Intelligibility Test based on Word Familiarity (FW03)”, NII Speech Resources Consortium, 2006.

本研究に関する研究業績

国際会議

- Yuxuan Du, Masato Akagi, “Speech Recognition in noisy conditions based on speech separation using Non-negative Matrix Factorization”, Proc. 2014 RISP International Workshop on Nonliner Circuits, Communications and Signal Processing, (to appear).