# 公立はこだて未来大学2025年度システム情報科学実習
# グループ報告書

**Future University Hakodate 2025 Systems Information Science Practice**

**Group Report**

**プロジェクト番号 / Project No.**
7

**プロジェクト名**
Dynamics Insights: 複雑パターンの解明

**Project Name**
Dynamics Insights: Unveiling Complex Patterns

**グループ名 / Group Name**
音班 / Sound group

**プロジェクトリーダー / Project Leader**
白木慶汰 / Keita Shiraki

**グループリーダー/ Group Leader**
服部夢叶 / Yuuki Hattori

**グループメンバー / Group Member**
服部夢叶 / Yuuki Hattori
坂井晴哉 / Haruya Sakai

**指導教員 / Advisor**
加藤譲 / Yuzuru Kato　義永那津人 / Natsuhiko Yosinaga　栗川知己 / Tomoki Kurikawa
ヴラジミールリアボフ / Volodymyr Riabov　リヴァーズダミアン / Damian Rivers

**提出日**
2026 年 1 月 21 日
**Date of Submission**
January 21,2026

# 概要

複数の音声が同時に存在する状況においては, 音声の聞き取りが困難になることが多い. このような問題は, オンライン会話や騒音環境下で特に顕在化しており, 円滑なコミュニケーションを妨げる要因となっている. これに対する有効な解決手法の一つとして音源分離が挙げられるが, 音源分離を実現するためには, まず各話者の音声的特徴を適切に捉え, 識別できることが重要である.

本プロジェクトでは, 同時発話音声における話者特定, およびノイズが存在する環境下における話者特定に取り組んだ. 音声信号から音響特徴量を抽出し, それらを機械学習アルゴリズムにより分類することで, 話者特定手法の有効性を評価した. 同時発話音声に対しては VAE を用いた潜在表現の学習を行い, ノイズ環境下では SVM を用いて話者分類を行った. 分析には JVSCorpus を使用した. その結果, 同時発話音声においては, 複数話者の中から発話者を約 9 割の精度で特定可能であり, ノイズ環境下では 7 割程度の特定性能であった. この成果は，音源分離技術に向けた前段階として有効である.

**キーワード**: 機械学習, 音声信号処理, 時系列解析

(※文責: 服部夢叶)

# Abstract

In situations where multiple speech signals are present simultaneously, speech perception often becomes difficult. This problem is particularly pronounced in online conversations and noisy environments, where overlapping speech interferes with smooth communication. One effective approach to addressing this issue is source separation; however, achieving reliable source separation requires accurately capturing and distinguishing the acoustic characteristics of each speaker. In this project, we investigated speaker identification in simultaneous speech as well as speaker identification under noisy conditions. Acoustic features were extracted from speech signals and classified using machine learning algorithms to evaluate the effectiveness of speaker identification methods. For simultaneous speech, a Variational Autoencoder (VAE) was employed to learn latent representations of speakers, while a Support Vector Machine (SVM) was used for speaker classification in noisy environments. The JVSCorpus dataset was used for the analysis. As a result, we confirmed that, in simultaneous speech scenarios, speakers could be identified from multiple candidates with an accuracy of approximately 90%.In noisy environments, a certain level of speaker identification performance was also achieved. These results indicate that the proposed approach is effective as a preliminary step toward source separation and has the potential to contribute to the development of robust speech processing techniques in complex acoustic environments.



**Keywords**: Machine learning, Voice Signal Processing, time series analysis


(※文責: 服部夢叶)

# Contents

# 1　はじめに

## 1.1　背景

日常生活において, 多人数の会話を聞く機会は多い. 対面の会話では, 発話タイミングが重なることは少ない. これは, うなずきなどの身体動作がターンテイキングと密接に関連しているためである [1]. 一方で, このような非言語的手がかりが利用できない環境では, 発話タイミングの重なりが増加することが知られている. 近年, コロナ禍の影響によりオンライン会話の利用が急速に拡大し [2], 今後もオンライン環境におけるコミュニケーション機会が増えることが予想される. オンライン会話では非言語情報が制限されるため発話タイミングが重なりやすく, 結果として音声の聞き取りが困難になる問題が生じる. これらの問題はオンライン会議に限らず, 日常生活においても環境雑音により会話が困難になる場面がある. この問題の一般的な解決手法として音源分離が挙げられる. しかし, 音源分離を効果的に行うためには, まず各話者の音声的特徴を適切に抽出する必要がある. 本プロジェクトは, この点に着目する.

(※文責: 服部夢叶)

## 1.2　目的

本プロジェクトの目的は, 同時発話音声およびノイズが存在する環境下において話者を特定することである. この問題を解決するため, 本研究では音声信号から音響的特徴量を抽出し, それらを入力とする機械学習アルゴリズムによる分類性能を指標として評価する.

本プロジェクトを通じて期待される効果として, 雑音などにより音声が聞き取りにくい状況においても, 目的とする音声信号を抽出可能な技術の発展につながる点が挙げられる. これは, 音源分離をはじめとする一般的な信号処理分野における任意信号抽出技術の向上にも寄与すると考えられる.

(※文責: 坂井晴哉・服部夢叶)

## 1.3　先行研究

先行研究として, 目的話者の音声の特徴を示す手がかりに基づき, 混ざった音声の中から目的話者の音声のみを抽出する技術 SpeakerBeam がある [4]. SpeakerBeam とは, 録音された目的話者の音声からその声の特徴量を抽出する話者特徴抽出ニューラルネットワーク (NN) と, 抽出した特徴量を補助入力として混合音声から目的話者の音声を抽出する目的話者抽出 NN, の 2 つの NN によって構成されている.

この研究では, 複数人の話者の音声データを録音して入力とし, 入力から特定の話者の特徴抽出を行い, 特定の一人の話者の音声を抽出している. 結果として, 目的の声が雑音や他人の声に対して, どれだけ大きくクリアに分離出来ているかを表している比率は平均して 8dB 以上であり, 高い抽出性能を達成していることが確認できる.

(※文責: 坂井晴哉)

# 2 関連研究

## 2.1 信号処理

- **ソース・フィルタモデル**：音響信号は, 音源がフィルタを通過した結果であるとするモデル. 人間の声では声帯が音源, 声道がフィルタと考える.[3]
  - $A(t)$ は声帯による励起信号（source）
  - $G(t)$ は声道のインパルス応答（filter）
  - $*$ は畳み込み演算を表す

$$S(t) = A(t) * G(t)$$

  - $S(\omega)$ は音声信号のスペクトル
  - $A(\omega)$ は励起信号のスペクトル
  - $G(\omega)$ は声道の伝達関数

  とすると

$$S(\omega) = A(\omega) \cdot G(\omega)$$

  で表される.

- **DFT（離散フーリエ変換）**：振幅情報だけでは不十分なため, 周波数特性を得るために用いる. 離散時間信号 $x[n]$ に対する DFT は、以下の式で定義される：

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j\frac{2\pi}{N}kn}, \quad k = 0, 1, \ldots, N-1$$

  逆離散フーリエ変換（IDFT: Inverse DFT）は、次のように定義される：

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] \cdot e^{j\frac{2\pi}{N}kn}, \quad n = 0, 1, \ldots, N-1$$

  ここで,

  - $x[n]$ は時系列信号（時間領域）
  - $X[k]$ は周波数成分（周波数領域）
  - $N$ は信号の長さ（点数）
  - $j$ は虚数単位

- **ケプストラム**：スペクトルの対数を逆フーリエ変換.

  音源とフィルタを和で表現可能になる. 人間の声の場合, ケプストラムの低次成分 (低周波成分) は. 声道フィルタ（口腔・咽頭などの共鳴特性）に対応する.

  対数スペクトルの分離

- $A[k]$ はスペクトル包絡の周波数特性
- $G[k]$ は音源の周波数特性とすると

$$X(k) = A(k) \cdot G(k)$$

対数を取ることで, 次のように分解できる：

$$\ln |X(k)| = \ln |A(k)| + \ln |G(k)|$$

ケプストラムの定義
実数信号 $x[n]$ に対するケプストラム $c[n]$ は, 以下のように定義される.：

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} \ln \left( |X[k]| \right) \cdot e^{j \frac{2\pi}{N} kn}$$

ここで,

- $X[k]$ は周波数スペクトル
- $|X[k]|$ は振幅スペクトルの大きさ (絶対値)

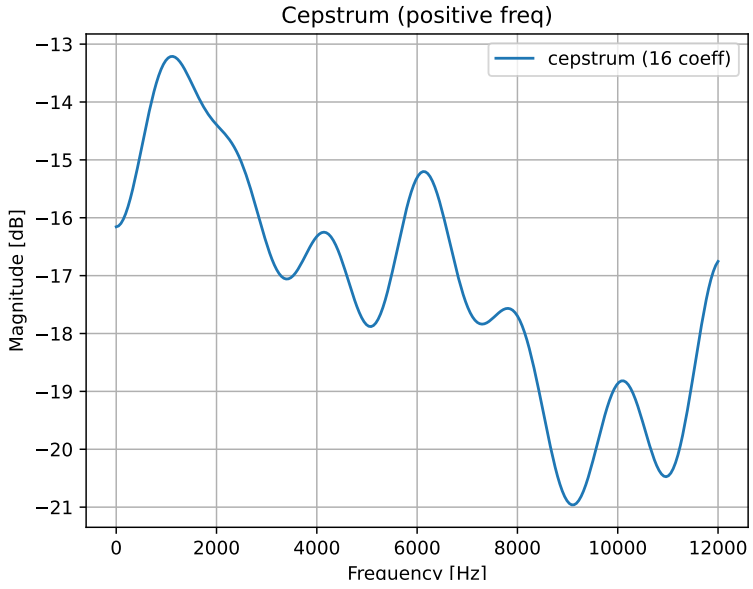


Figure 1: ケプストラム

- **メルフィルタ**：人間の聴覚特性に基づく周波数スケーリングを行い, より有効な特徴抽出を可能にする.

## 基準値を用いたメルスケール変換

基準値として以下を用いる.：

$$f_0 = 700 \quad [\text{Hz}], \quad m_0 = 2595 \quad [\text{mel}]$$

### 1. メル変換式

$$m(f) = m_0 \cdot \log_{10}\left(1 + \frac{f}{f_0}\right)$$

### 2. 逆変換式（mel → Hz）

$$f(m) = f_0 \cdot \left(10^{\frac{m}{m_0}} - 1\right)$$

- **LPC（線形予測符号化）**：将来の値をそれまでの標本群の線型写像として予測する方法を用いた符号化
離散時間音声信号 $x[n]$ は，以下のように予測される：

$$\hat{x}[n] = \sum_{k=1}^{p} a_k \, x[n-k]$$

ここで，$p$ は予測次数，$a_k$ は線形予測係数である.
予測誤差（残差信号）は次式で定義される：

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{k=1}^{p} a_k \, x[n-k]$$

線形予測係数 $a_k$ は，平均二乗誤差

$$E = \sum_{n} e[n]^2$$

を最小化するように求められる.
この最小化問題は，Yule–Walker 方程式として次のように表される：

$$\sum_{k=1}^{p} a_k \, R[|i-k|] = R[i], \quad i = 1, 2, \ldots, p$$

ただし，$R[i]$ は自己相関関数であり，

$$R[i] = \sum_{n} x[n] \, x[n-i]$$

で定義される.

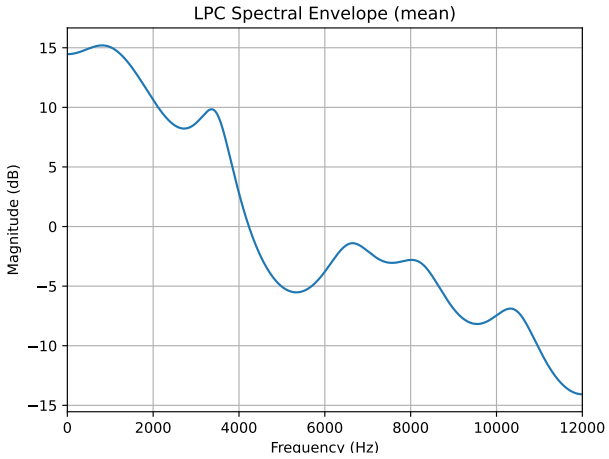


Figure 2: LPC

(※文責: 服部夢叶)

## 2.2 VAE

VAE（Variational Autoencoder）は確率的生成モデルの一種であり，エンコーダ，潜在変数からなる潜在空間，およびデコーダによって構成される．潜在空間は入力データの本質的な特徴を低次元で表現しており，潜在変数を操作することでデータの生成や補間が可能となる．VAE（Variational Autoencoder）は，潜在変数 $z$ を導入した確率的生成モデルである．観測変数 $x$ に対する周辺尤度は次式で表される：

$$p(x) = \int p(x \mid z)p(z)\,dz$$

しかし，この積分は一般に解析的に求めることが困難であるため，真の事後分布 $p(z \mid x)$ を $q_\phi(z \mid x)$ により近似する．

VAE では，対数尤度 $\log p(x)$ の下界である ELBO(Evidence Lower Bound）を最大化する：

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - D_{\mathrm{KL}}\left(q_\phi(z \mid x) \,\|\, p(z)\right)$$

ここで，

- $q_\phi(z \mid x)$：エンコーダ（近似事後分布）
- $p_\theta(x \mid z)$：デコーダ（生成分布）
- $p(z)$：事前分布（通常は標準正規分布）
- $D_{\mathrm{KL}}(\cdot\|\cdot)$：Kullback–Leibler ダイバージェンス潜在変数 $z$ は，再パラメータ化トリックにより次のように表される：

  $$z = \mu + \sigma \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$$

  損失関数として次を用いた
  $\mathrm{L} = \text{-}\ \mathrm{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + D_{\mathrm{KL}}\left(q_\phi(z \mid x) \,\|\, p(z)\right)$

(※文責: 服部夢叶)

## 2.3 SVM(サポートベクターマシン)

SVM とは, 空間内の各クラス間の距離を最大化する最適な超平面を見つけることでデータを分類する, 教師あり機械学習アルゴリズムである. クラスを区別する複数の超平面が見つかるため, 点間の境界までの最小距離 (マージン）を最大化することで, アルゴリズムはクラス間の最適な決定境界を見つけることができる. これは、新しいデータに対して良好に一般化し, 正確な分類予測を行うことを可能にする.

(※文責: 坂井晴哉)

# 3　技術詳細・データセット

本研究の実行環境および使用データセットを以下に示す.

- 実行環境：Google Colaboratory
- 使用言語：Python
  - 音声データセット：JVSCorpus
  - 一人につき共通の 100 セリフ
  - 総計 100 人の音声データあり
  - 各音声はおおよそ 10 秒程度
- 雑音データセット：ESC-50-master
- 音声形式：モノラル
- サンプリングレート：24,000 Hz

VAE に入力する特徴量として，以下の音響特徴を用いた。

- LPC 係数の平均値:16 次元
- 実部の振幅スペクトルの平均値:513 次元
- 実部の振幅スペクトルの標準偏差:513 次元

SVM に入力する特徴量として，以下の音響特徴を用いた。

- mel-log スペクトル:64 次元
- MFCC(メルフィルタスペクトラム係数):30 次元

(※文責: 服部夢叶)

# 4　実験内容

## 4.1　同時発話音声からの話者特定

対象とする話者は 10 名とし,2 人の同時発話音声に含まれる話者の特定を行った. 同時発話音声は, 単独話者音声の波形を線形和として合成することで生成した. この性質を踏まえ, 線形性を保つ特徴量および線形活性化関数のみを用いた VAE を設計した. 特徴量はスペクトルの実部の時間平均と LPC の時間平均である.

学習段階では, 単独話者の音声から得られた特徴量を VAE に入力し, 潜在空間における潜在変数を獲得した. さらに,VAE の潜在空間からサンプリングされた潜在変数を用いて, ロジスティック回帰モデルによる話者識別器を学習した.

推定段階では, 同時発話音声から得られた特徴量を学習済み VAE に入力し, 対応する潜在変数を導出した. 得られた潜在変数を, あらかじめ学習したロジスティック回帰モデルに入力することで, 同時発話音声中に含まれる話者の分類および特定を行った.

(※文責: 服部夢叶)

## 4.2　ノイズ環境下における話者特定

ノイズが存在する環境下における音声の学習性能を評価するため, 音声に環境音を加えた複合音声を対象とし，10 人の中で 1 人の発話者を特定する話者分類を行った. 分類器には SVM を用い，ノイズが混在する条件下における音声特徴の学習および識別能力を検証した.

特徴量として，メルスペクトルの対数の時間平均および MFCC（Mel-Frequency Cepstral Coefficients）を用いた．これらの特徴量を入力として SVM を学習し，ノイズ環境下において音声情報がどの程度安定して学習・識別されるかを評価した.

(※文責: 坂井晴哉)

# 5　結果

## 5.1　同時発話音声からの話者特定

Figure 3 は,10 名の話者から任意の 2 名が同時に発話した場合における話者特定精度を示したものである．横軸および縦軸はそれぞれ同時発話を行った話者を表し, 各格子点の値は, 当該話者ペアに対する話者分類の精度を示している. 線形な特徴量および線形な活性化関数を用いたモデルにおいて,10000 回の試行を行った結果, 同時発話している 2 名の話者を 10 名の話者集合の中から約 89 % の精度で特定できることが確認された.

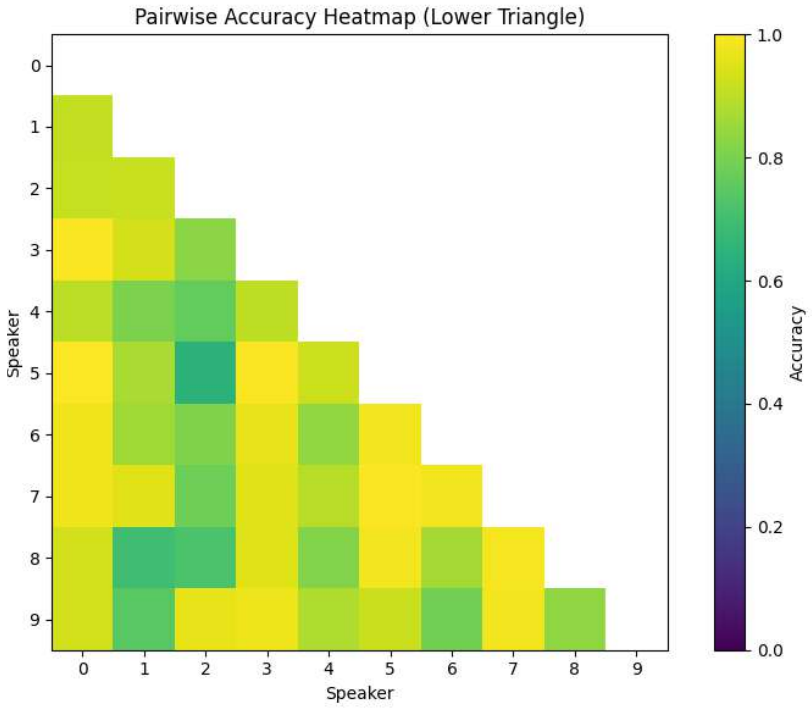


Figure 3: VAE の結果

(※文責: 服部夢叶)

## 5.2　ノイズ環境下における話者特定

Figure 4 は 10 人の中で 1 人の発話者を特定する話者分類の結果を混同行列で表したものである. 混同行列とは, 分類モデルの性能を評価する表のことで, 左斜めに対角線状になっているマスが正しく分類されていて, 他のマスは誤って分類されている. 特定精度は,6 割程度である.

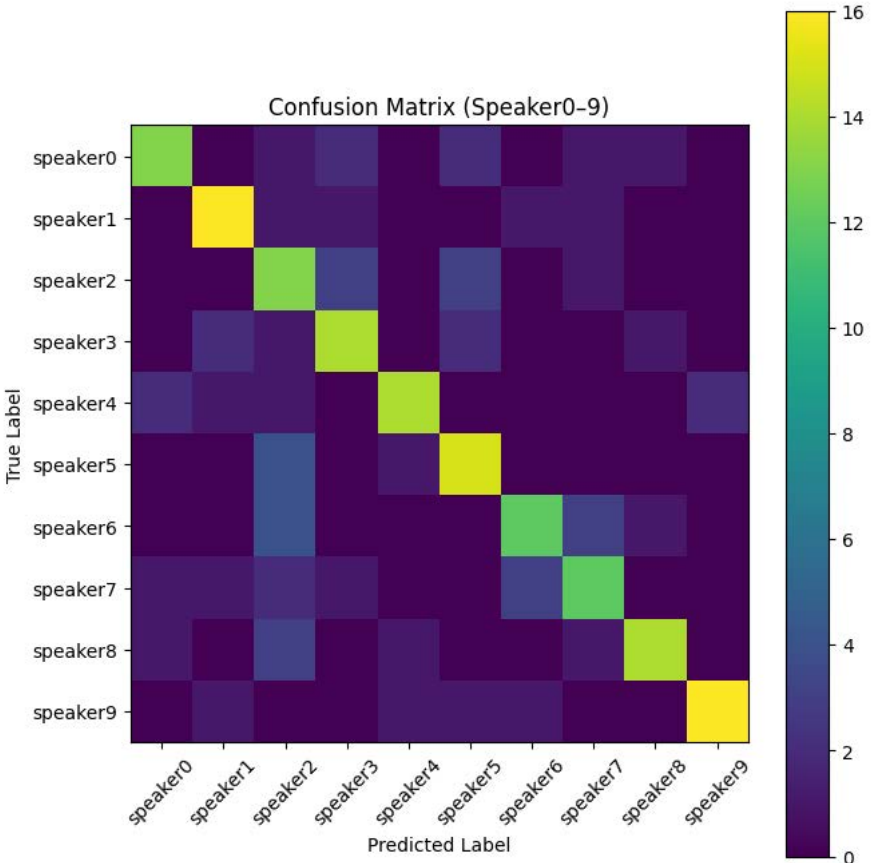


Figure 4: SVM の結果

(※文責: 坂井晴哉)

# 6　考察

## 6.1　同時発話音声からの話者特定

同時発話音声が各話者の発話波形の線形和として表現できるという仮定に基づき, 線形な特徴量および線形活性化関数のみを用いた VAE を設計した. この設計により, 単独話者音声のみを用いて学習した潜在空間においても, 同時発話音声が話者ごとの特徴量の線形和として表現されることが期待される. 実際に, 単独話者音声から学習した潜在空間を用いて,2 人の同時発話音声に含まれる話者を分類可能であることが確認された. この結果は同時発話音声に対しても単独話者音声との対応関係を保った表現が可能であることを示している. 一方で, 本手法は線形な特徴量および活性化関数のみを用いているため, モデルの表現力が制限されるという課題がある. その結果, 学習話者数が増加した場合や話者間の声質差が小さい場合には, 潜在空間上で話者を十分に分離できず, 分類性能が低下する. さらに,3 人以上の同時発話音声を対象とした場合に分類性能が低下したことは, 本手法の線形性に起因すると考えられる. また, 非線形な特徴量や活性化関数を用いた場合, 波形の線形和から導出される特徴量は加法性を失う. その結果, 単独話者音声から構築された潜在空間との対応関係が崩れ, 同時発話音声中の話者を正確に特定することができなかった.

同時発話者が 2 人であれば, 特定精度が高いことから, 本手法は音源分離などの事前処データとして活用できる可能性がある.

(※文責: 服部夢叶)

## 6.2　ノイズ環境下における話者特定

speaker1,9 は他の話者と特徴が被りにくく,MFCC で特徴が明確に抽出されていて性能が良い. しかし,speaker6,7 は誤分類が多く, 性能が悪い部分も見受けられた. このことから, 本手法は話者固有の特徴をある程度捉えられているものの, 話者間の分離性能には限界があるといえる.

今後は, 特徴量の拡張や正規化の工夫, より表現力を持つモデルの導入, データの拡張などを行うことで, 話者間の識別性能向上が期待されると考える.

(※文責: 坂井晴哉)

# 7　参考文献


1. 横山 真男, 青山 一美, 菊池 英明, 白井 克彦: 人間とロボットのコミュニケーションにおける非言語情報の利用, 情報処理学会研究報告音声言語処理巻 1998 号 49(1998-SLP-021)p69-74(1998)

2. 濱野 和佳, 後藤 学: コロナ禍におけるオンラインコミュニケーションツールの利用状況と利用者の受け止め,INSS JOURNAL Vol. 28 204,226(2021)

3. 田中 聡久, 川村新「音声音響信号処理の基礎と実践」コロナ社 2021

4. 聞きたい人の声に耳を傾ける AI ——深層学習に基づく音声の選択的聴取技術 SpeakerBeam https://journal.ntt.co.jp/article/14481 (2026/1/14 アクセス)