音声の意味的類似度に基づいた対照学習によるリアルタイム応答選択

大中緋慧 [1,2] 大西一誉 [1,2] 吉野幸一郎 [1,2,3]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所ガーディアンロボットプロジェクト [3] 東京科学大学
{onaka.hien.oj5,onishi.kazuyo.oi5,koichiro}@naist.ac.jp

概要

応答タイミングや相槌などは音声対話システムの表現力に関係する重要な要素である．この背景に基づいて，我々は以前の研究で応答タイミングと遅延緩和のための短文応答を同時予測するモデルを提案した．このモデルは，対話文脈に基づく適切なタイミングを決定するとともに，適切な短文応答も合わせて選択する．本論文では，音声の客観評価指標に基づく重み・順序付き対照学習を提案し，短文応答選択を改善する．客観評価スコアに基づく学習により，ランキング付けにおける一貫性の改善を実現することを目指す．評価実験では，提案手法が短文応答選択における top-k % を改善することを示した．

1 はじめに

人間は対話の中で，スムーズな情報のやり取りを実現するためにターンテイキングを行う．本稿でのターンテイキングは，発話権を調整する振る舞いの総称を指す．ターンテイキングのモデル化は対話システムにとっても自然な対話を実現するための重要な課題である．古典的なターンテイキングモデル [1, 2] は，無音区間に基づいて高精度にユーザ発話の終端検出を行うことを目的としていた [3]．

こうした古典モデルはユーザ発話の終了後にシステム発話を開始するため，対話破綻を発生させないという点で優れている．しかし，実際の対話においてはバックチャネルを含む多様な応答タイミングが，より円滑な対話の実現に重要な役割を果たしているため，これらの振る舞いもモデリングできることが望ましい．深層学習 [4, 5] や大規模言語モデル (Large Language Model: LLM) [6, 7] の発展により，自然な音声合成 [8, 9] や応答生成が可能となった．それに伴って，バックチャネルや応答タイミングなどの研究も盛んになってきている [10, 11]．この研究は二つの方向性に大きく分けることができる．

一つは，対話システム全体を End-to-end で最適化することで暗黙的にターンテイキングの振る舞いを獲得するものである [12, 13]．この枠組みの一つである Moshi [13] は，二話者の音声ストリームを逐次的に生成するようにモデルを学習する．その結果，学習データに含まれるターンテイキング的な振る舞いを暗黙的に獲得する．推論時には，いずれか一方のストリームをシステム音声として扱う．このアプローチは，ターン制対話の仮定を取り払い，優れた応答速度や自然な相槌などを実現できる．一方で，モデルの訓練には数万時間以上のデータが必要であり，制御性も低いという問題を抱えている．

もう一つのアプローチは，カスケード対話システム [14] の中にストリーミング処理可能なターンテイキングモデルを導入するものである．代表的な手法の一つとして，Voice activity projection (VAP) [15, 16] が存在する．このモデルは，二話者の音声を逐次的に受け取り，二秒先の音声活動を予測するように分類タスクで学習される．推論時には，予測された音声活動に基づいて逐次的にシステムのバックチャネルやメイン応答のタイミングを決定できる．このようなアプローチは比較的少ないデータでモデルを構築でき，対話システム内の各モジュールをカスタマイズ可能で制御性が高いという点で優れている．

しかしながら，カスケード対話システム上でのターンテイキングモデルでは，生成時の遅延が大きな問題となる．モデルが適切な応答タイミングを決定できたとしても，カスケードモデルの構成要素である対話応答生成と音声合成が律速となる．この課題を解決する方法として，フィラーを利用した遅延緩和のアプローチがよく知られている [17, 18]．これは，生成遅延を待機する間に，事前に用意した

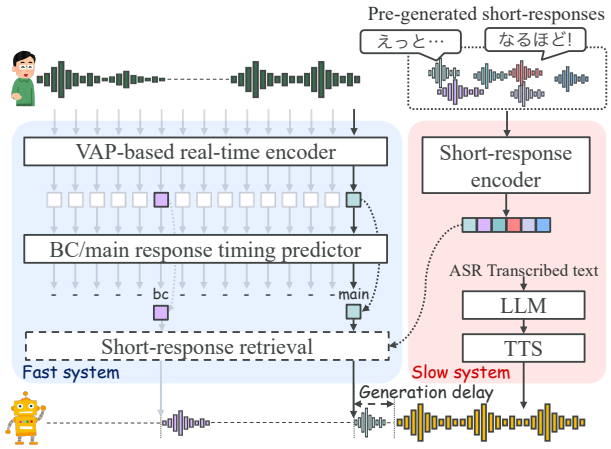

図 1　本研究におけるフレームワーク図．応答タイミングを逐次的に予測する早いシステムと，短文応答集合を一度だけエンコードする遅いシステムで構成される．

フィラー音声を再生することでユーザ視点での遅延を解消するものである．このアプローチを踏まえて，我々は短文応答選択と応答タイミング予測の同時予測モデルを提案した [19]．この手法は，VAP モデルに基づいて対話情報から逐次的に応答タイミングを予測しつつ，短文応答との対照学習を適用することで，文脈にあった応答タイミングと短文応答を同時に選択して遅延緩和に用いるものである．

本論文では，短文応答が持つ意味役割に着目し，音声客観評価指標に基づく順序・重み付き対照学習を提案する．提案手法は次の二ステップから構成される．まず，SpeechBERTScore [20] を用いて，二つの短文応答ペア同士の意味的な類似度を与える．学習時には，この類似度に基づいて二つの損失項を導入する．第一に，InfoNCE [21] を重み付きに拡張した損失関数を導入し，区別の難しい細かな違いに着目した学習を可能にする．第二に，triplet margin loss [22] に順序の制約を導入したペナルティ損失を追加する．これは，類似度に基づくランキングと予測結果が一致するような制約を与えるものであり，隣接する応答同士の意味的な一貫性の改善が期待される．評価実験では，提案された二つの損失関数により応答タイミング予測の精度を低下させることなく，短文応答選択における top-k % を大幅に改善できることを示した．

2　問題設定

本研究の問題設定について述べる．本研究におけるモデルは，図 1 に示すように，タイミング予測と短文応答選択の二つのタスクから構成される．

2.1　VAP に基づく応答タイミング予測

応答タイミング予測は，VAP モデル [16] をベースとして実現される．まず初めに，チャンクに分割された二話者の音声を Contrastive Predictive Coding (CPC) encoder [23] に入力し，f_{vap} hz の時系列特徴を抽出する．それを self-attention と cross-attention の組み合わせからなる Transformer [5] で処理し，最後に二話者の特徴が統合される．モデルは，未来の音声活動ラベルに対するクロスエントロピー損失 $\mathcal{L}_{\text{vap}}$ [15] と，現在の音声活動ラベルを予測するクロスエントロピー損失 $\mathcal{L}_{\text{vad}}$ [15] で学習される．

本フレームワークにおける応答タイミング予測は，先行研究 [11] に基づいて，x フレーム先のシステムの音声活動を予測する三値のクロスエントロピー損失 [1] $\mathcal{L}_{\text{timing}}$ による事前学習済みモデルの fine-tuning で実現される．

2.2　リアルタイム短文応答選択

ここでは，リアルタイム短文応答選択の問題設定について述べる．我々のアプローチは，先行研究におけるオフライン動作する短文応答の対照学習 [24] をリアルタイムに拡張したものとして捉えられる．

まず，対話コンテキストとして，VAP encoder の隠れ特徴のうち，応答が立ち上がるタイミングのフレームのみを抽出し，Multi-layer perceptron (MLP) 層を通した特徴量 $\boldsymbol{h}_{\text{vap}} \in \mathbb{R}^D$ を得る．ここで D は対照学習における特徴次元を示す．次に，短文応答について Self-supervised learning (SSL) モデル [25] と MLP 層，プーリング層を通して得られる正例特徴 $\boldsymbol{h}^{+}_{\text{sr}} \in \mathbb{R}^D$ を得る．また，同一セッション同一話者の異なる短文応答集合からサンプリングされた負例特徴 $\boldsymbol{h}^{-}_{\text{sr}} \in \mathbb{R}^{(N+1)\times D}$ を得る．ここで，N は各バッチでの負例数を示し，$\boldsymbol{h}^{-}_{\text{sr},0} = \boldsymbol{h}^{+}_{\text{sr}}$ とする．$\boldsymbol{h}_{\text{vap}}$ をアンカーとして InfoNCE 損失 [21] を適用する．

$$\mathcal{L}_{\text{nce}} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}^{+}_{\text{sr}})/\tau)}{\sum_{n=0}^{N} \exp(\text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}^{-}_{\text{sr},n})/\tau)}, \tag{1}$$

$$\text{sim}(\boldsymbol{a}, \boldsymbol{b}) = (\boldsymbol{a} \cdot \boldsymbol{b})/(||\boldsymbol{a}|| \cdot ||\boldsymbol{b}||). \tag{2}$$

ここで，τ は温度パラメータを示す．推論時には，応答タイミングの立ち上がり時の特徴 $\hat{\boldsymbol{h}}_{\text{vap}}$ に合致する短文応答が選択されることが期待される．

3　提案手法

本章では，提案手法について述べる．

1)　'none' (応答なし)，バックチャネル，メイン応答の三つ．

3.1 基本アイディア

2.2 節で述べた短文応答選択手法では，チャンスレートを比較手法として top-k % 客観評価で優れたスコアが得られることを確認した [19]．ただし分析の結果，ランキング付けの中で隣り合うランクの短文応答同士の意味的な一貫性が低いことも確認された．短文応答は対話行為的な意味役割を持つため，学習データ中の正例に類似する意味役割を持つ短文応答もランクの上位に位置付けるように学習されることが望ましい．その上で，類似した短文応答同士を正確に区別できることが理想的である．本稿では，これら二つの課題に対応するために，音声客観指標である SpeechBERTScore [20] に基づく順序・重み付き対照学習を提案する．

3.2 SpeechBERTScore とその活用

SpeechBERTScore [20] は Saeki らによって提案された音声品質の客観評価指標である．この指標は，評価したい音声と参照音声の両方からの SSL 連続値特徴に対して BERTScore [26] を計算し，その意味的な一致性を音声 SSL 空間上で捉えるものである．従来の音声の自動評価指標 [27, 28] とは異なり，両音声間の長さが異なっていても適用できることが主要な利点である．評価対象の音声波形から抽出された SSL 特徴量を $\hat{\boldsymbol{z}} \in \mathbb{R}^{T_{\text{eval}} \times D}$，参照音声から抽出された SSL 特徴量を $\boldsymbol{z} \in \mathbb{R}^{T_{\text{ref}} \times D}$ とする．$T_{\text{eval}}, T_{\text{ref}}, D$ はそれぞれ評価音声，対照音声の SSL 特徴のフレーム長，SSL モデルの特徴次元を示す．ここで，評価対象の音声の SpeechBERTScore $\mathcal{S}(\cdot,\cdot)$ は次の式で定義される．

$$\mathcal{S}(\hat{\boldsymbol{z}}, \boldsymbol{z}) = \frac{1}{T_{\text{eval}}} \sum_{i=1}^{T_{\text{eval}}} \max_{j} \text{sim}(\hat{\boldsymbol{z}}_i, \boldsymbol{z}_j). \tag{3}$$

上述の評価指標は合成音声，劣化音声での人間の主観評価と高く相関することが実験で示されている．

本研究では，同指標の音声上での意味的な一致を捉える性質に着目し，本来の用途とは異なるものの，短文応答同士の類似度の近似値として利用することを考える．正例の短文応答の SSL 特徴 $\boldsymbol{z}^{+} \in \mathbb{R}^{T_p \times D}$，負例群の SSL 特徴 $\boldsymbol{z}^{-} = \{\boldsymbol{z}_n^{-}\}_{n=1}^{N}$ として，類似度 $\boldsymbol{s} \in \mathbb{R}^{N}$ は $s_n = \mathcal{S}(\boldsymbol{z}_n^{-}, \boldsymbol{z}^{+})$ で計算される．

3.3 重み付き InfoNCE

式 (1) を 3.2 節で述べたスコアを用いて拡張することを考える．ハードネガティブサンプリング [29] の考え方に従い，区別の難しい類似度の高い負例に大きな重みを割り振る損失関数を定義する．

$$\mathcal{L}_{\text{w-nce}} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}_{\text{sr}}^{+})/\tau)}{\sum_{n=0}^{N} w_n \exp(\text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}_{\text{sr},n}^{-})/\tau)} \tag{4}$$

$$w_n = \frac{\exp(s_n/\tau_w)}{\sum_{i=1}^{N} \exp(s_i/\tau_w)} \tag{5}$$

ここで，τ_w は重み計算時の温度パラメータを示し，$w_0 = 1.0$ とする．温度パラメータが小さいほど一サンプルに重みが集中し，$\tau_w \to +\infty$ で式 (1) と等価となる．この損失関数を用いることで，負例群の中で正例によく似たサンプルが強く影響するため，より細かな違いを区別するような学習が期待される．

3.4 ペナルティ損失

3.3 章で述べた重み付き損失は，細かな違いを捉える学習として効果的であるが，特定のサンプルに重みが集中することで大域的な判別精度を悪化させる可能性がある．この問題を緩和し，類似度の大きさに従うランキングを促進するために triplet loss [22] に基づくペナルティ損失を導入する．損失項は，$s_1 \geq \ldots s_j \geq \ldots s_J$ となるような順序でソートされた短文応答の特徴 $\boldsymbol{h}_{\text{sr}}^{-}$ に対して，次の拡張された triplet loss を用いて計算される．

$$\mathcal{L}_{\text{penalty}} = \sum_{i=0}^{N-1} \sum_{j=i+1}^{N} \text{ReLU}(\text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}_{\text{sr},j}^{-}) - \text{sim}(\boldsymbol{h}_{\text{vap}}, \boldsymbol{h}_{\text{sr},i}^{-}) + \alpha) \tag{6}$$

ここで，α はマージンを示しており，$\boldsymbol{h}_{\text{sr},0}^{-} = \boldsymbol{h}_{\text{sr}}^{+}$ である．この損失は，意味的類似度に基づくランキングに反するコサイン類似度のランキングが生じた場合に，それを抑制するものである．この機能により，最終的に獲得されたモデルは正例とよく似た負例を高いランクへ，大きく異なる負例を低いランクへと分類する空間を獲得することが期待される．

最終的な損失関数は $\mathcal{L}_{\text{timing}}, \mathcal{L}_{\text{w-nce}}, \mathcal{L}_{\text{penalty}}$ と補助損失 $\mathcal{L}_{\text{vap}}, \mathcal{L}_{\text{vad}}$ の重み付き和で表される．

4 実験条件

4.1 学習データ

学習及び評価データとして，高齢者傾聴対話コーパス [30] を使用した．一方の参加者はナビゲータ，もう一方の高齢の参加者はユーザである．ナビゲータがニュース情報の提供，ユーザがそれに対する意見を述べ，ナビゲータが傾聴するという手順を何度

か繰り返した 20–30 分の対話が 60 収録されたものである．同コーパスは自発音声やフィラーなどを多分に含むという点に着目して採用した．

タイミングラベル抽出: コーパスに含まれる各話者の発話区間ラベルを利用して，バックチャネルとメイン応答のラベリングを行った．話者が遷移する際の発話区間ラベルに基づいて，遷移前の話者の発話終了から 1 秒以内に遷移後の話者の発話が終了する場合をバックチャネル，それ以外の場合をメイン応答として自動的にラベルが付けられた．

短文応答抽出: より細かい粒度の音声区間を抽出するために，バックチャネル，メイン応答のそれぞれに対して，追加で silero-vad [31] が適用された．その後，2 秒以上になるまでチャンクを前方から結合し，得られた音声を短文応答として収集した．各セッションの各話者で，それぞれ 200 前後の短文応答が得られた．学習時には，同一セッション同一話者の集合からランダムに負例がサンプリングされた．

4.2 モデル設定

ベースモデルとして，MaAI [32] で公開されている ‘vap_mc_state_dict_jp_20hz_10000msec‘ [33] を使用した．短文応答エンコーダは，日本語で学習された Hubert base [34] を使用し，最初の 10 層のパラメータは凍結した．最終層の出力に平均プーリングを適用し，その後 2 層 MLP を用いることで最終的な特徴量が抽出された．最終的な特徴量次元は 128 であった．3.2 節で述べた SpeechBERTScore の計算には，日本語で学習された Hubert large [35] の 12 層目特徴量を用いた．温度パラメータ τ_w は 0.05 とした．

モデルは 50 エポックの反復で学習された．最適化アルゴリズムには，AdamW [36] が用いられ最初の 10 エポックで $5e$-5 となるように warmup が適用された．その後 50 エポックまで，$1e$-6 に向けてコサイン関数に従って学習率が減衰された．損失関数の重みは $w_{\text{timing}} = 1.0$, $w_{\text{w-nce}} = 5.0$, $w_{\text{penalty}} = 5.0$, $w_{\text{vap}} = 0.2$, $w_{\text{vad}} = 0.05$ であった．$\mathcal{L}_{\text{penalty}}$ は 10 エポック目から導入され 30 エポックまで warmup が適用された．

5 実験結果

正例がコサイン類似度に基づくランキング付けにおいてどの位置に属するかを評価する top-k % を用いて評価を行った．ベースラインとして，チャンスレートである Random, 通常の InfoNCE 損失を使用する Baseline とそれに $\mathcal{L}_{\text{penalty}}$ を加えたもの, 提案手法から提案する損失をそれぞれ除いたものを用いた．

表 1 提案手法の top-k % による客観評価

Methods	Top-k %				
	1	5	10	25	50
Random	1.0	5.0	10.0	25.0	50.0
Baseline ($\mathcal{L}_{\text{nce}}$)	9.4	24.5	37.1	61.4	81.4
+ $\mathcal{L}_{\text{penalty}}$	10.7	24.9	35.8	55.0	73.1
Proposed (only $\mathcal{L}_{\text{w-nce}}$)	12.0	28.4	**46.3**	**67.8**	**88.5**
Proposed (only $\mathcal{L}_{\text{penalty}}$)	11.4	23.8	32.3	50.2	69.4
Proposed	**17.5**	**31.4**	42.4	59.8	75.1

表 2 タイミング推定タスクにおける客観評価

Methods	F1 [%]	Precision [%]	Recall [%]
Baseline ($\mathcal{L}_{\text{nce}}$)	85.6	89.7	82.5
Proposed	86.0	90.0	83.1

評価結果を表 1 に示す．表から，top-1, 5, 10 の指標で提案手法が優れたスコアを示すことが分かる．また，二つの損失関数の組み合わせが効果的であることが確認された．一方で，top-25, 50 では $\mathcal{L}_{\text{penalty}}$ の導入がスコアを低下させることが確認された．この現象は，ランキングの一貫性が強化された結果発生したと考えられる．実際にはある対話コンテキストに対して，ポジティブ・ネガティブな応答のどちらを行うかは応答側コンテキストなしでは一意に決定しない．テストセットの正解とのミスマッチが発生した場合に，誤ったサンプルとの一貫性の強い $\mathcal{L}_{\text{penalty}}$ ありの手法では，本来の正解であるサンプルがより低いランクへと押し出されている可能性がある．この問題を解決することは今後の課題である．

最後に，応答タイミング推定における F1, Recall, Precision を表 2 に示す．導入された損失項がタイミング推定タスクに悪影響を与えないことがわかる．

6 まとめ

本研究では，リアルタイム音声対話のための短文応答選択における対照学習の改良手法を提案した．提案手法では，音声の意味的類似度に基づいて，負例の重要度の重み・順序付けを行う．得られたスコアを利用した対照学習により，細かな違いを捉えるような特徴空間の獲得や，ランキング付けにおける隣接ランク同士の応答の一貫性の改善が期待される．今後は，実対話における自然性評価や，応答側の未来のコンテキストを取り入れることが可能な枠組みへの拡張などを進める予定である．

謝辞

本研究の一部は科研費 23K24910 の助成を受けて実施された.

参考文献

[1] R. Sato and R. Higashinaka et al. Learning decision trees to determine turn-taking by spoken dialogue systems. In **Proc. of ICSLP**, pp. 861–864, 2002.

[2] A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In **Proc. of NAACL**, pp. 629–637, 2009.

[3] N. G. Ward and A. G. Rivera et al. Root causes of lost time and user stress in a simple dialog system. In **Proc. of Interspeech**, pp. 1565–1568, 2005.

[4] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. In **Proc. of ICLR**, 2015.

[5] A. Vaswani and N. Shazeer et al. Attention is all you need. In **Proc. of NeurIPS**, pp. 5998–6008, 2017.

[6] OpenAI. GPT-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[7] H. Touvron and T. Lavril et al. LLaMA: Open and efficient foundation language models. **arXiv preprint arXiv:2302.13971**, 2023.

[8] Y. Wang and R. J. Skerry-Ryan et al. Tacotron: Towards end-to-end speech synthesis. In **Proc. of Interspeech**, pp. 4006–4010, 2017.

[9] J. Kim and J. Kong et al. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In **Proc. of ICML**, Vol. 139, pp. 5530–5540, 2021.

[10] M. Roddy and N. Harte. Neural generation of dialogue response timings. In **Proc. of ACL**, pp. 2442–2452, 2020.

[11] K. Inoue and D. Lala et al. Yeah, un, oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In **Proc. of NAACL (long)**, pp. 7171–7181, 2025.

[12] T. A. Nguyen and E. Kharitonov et al. Generative spoken dialogue language modeling. **Trans. Assoc. Comput. Linguistics**, Vol. 11, pp. 250–266, 2023.

[13] A. Défossez and L. Mazaré et al. Moshi: a speech-text foundation model for real-time dialogue. **arXiv preprint arXiv:2410.00037**, 2024.

[14] Y. Chiba and K. Mitsuda et al. The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models. In **Proc. IWSDS**, pp. 1–6, 2024.

[15] E. Ekstedt and G. Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In **Proc. of Interspeech**, pp. 5190–5194, 2022.

[16] K. Inoue and B. Jiang et al. Multilingual turn-taking prediction using voice activity projection. In **Proc. of LREC-COLING**, pp. 11873–11883, 2024.

[17] T. Shiwa and T. Kanda et al. How quickly should communication robots respond? In **Proc. of HRI**, pp. 153–160, 2008.

[18] N. Ohshima and K. Kimijima et al. A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings. In **Proc. of RO-MAN**, pp. 325–330, 2015.

[19] 大中緋慧, 河野誠也, 他. リアルタイム音声対話システムのための応答タイミングと短文応答の同時予測. In **Annual Meeting of NLP**, pp. 4063–4067, 2025.

[20] T. Saeki and S. Maiti et al. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging NLP evaluation metrics. In **Proc. of Interspeech**, 2024.

[21] A. Oord and Y. Li et al. Representation learning with contrastive predictive coding. **arXiv preprint arXiv:1807.03748**, 2018.

[22] J. Wang and Y. Song et al. Learning fine-grained image similarity with deep ranking. In **Proc. of CVPR**, pp. 1386–1393, 2014.

[23] M. Riviere and A. Joulin et al. Unsupervised pretraining transfers well across languages. In **Proc. of ICASSP**, pp. 7414–7418. IEEE, 2020.

[24] L. Qian and G. Skantze. Joint learning of context and feedback embeddings in spoken dialogue. In **Proc. of Interspeech**, pp. 2955–2959, 2024.

[25] W. Hsu and B. Bolte et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. **IEEE/ACM Trans. ASLP**, Vol. 29, pp. 3451–3460, 2021.

[26] T. Zhang and V. Kishore et al. BERTScore: Evaluating text generation with BERT. In **Proc. of ICLR**, 2020.

[27] A.W. Rix and J.G. Beerends et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In **Proc. of ICASSP**, pp. 749–752 vol.2, 2001.

[28] C. H. Taai and R. C. Hendriks et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In **Proc. of ICASSP**, pp. 4214–4217, 2010.

[29] M. Kaya and H. S. Bilge. Deep metric learning: A survey. **Symmetry**, Vol. 11, No. 9, p. 1066, 2019.

[30] K. Yoshino and H. Tanaka et al. Japanese dialogue corpus of information navigation and attentive listening annotated with extended ISO-24617-2 dialogue act tags. In **Proc. of LREC**, pp. 2922–2927, 2018.

[31] Silero Team. Silero VAD: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2024.

[32] MaAI-Kyoto. A real-time and light-weight software for generation of non-linguistic behaviors in conversational ais. https://github.com/MaAI-Kyoto/MaAI.

[33] K. Inoue and Y. Okafuji et al. A noise-robust turn-taking system for real-world dialogue robots: A field experiment. In **Proc. of IROS**, 2025.

[34] AIST Intelligent Media Processing Research Team. imprt/kushinada-hubert-base. https://huggingface.co/imprt/kushinada-hubert-base.

[35] AIST Intelligent Media Processing Research Team. imprt/kushinada-hubert-large. https://huggingface.co/imprt/kushinada-hubert-large.

[36] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In **Proc. of ICLR**, 2019.