Name Pronunciation Extraction and Reuse in Human-Robot Conversations

James Kennedy
james.kennedy@disney.com
Disney Research
Glendale, California, USA

Naveen Kumar
naveen.kumar@disney.com
Disney Research
Glendale, California, USA

Maike Paetzel-Prüsmann
maike.paetzelprusmann@disney.com
Disney Research
Zurich, Switzerland

ABSTRACT

Personalization in human-robot interaction (HRI) has been shown to have powerful effects on both users' perception of robots and objective interaction outcomes. Calling a human user by their name, an important signal to communicate understanding the user and memorizing information about them, remains an ongoing challenge in HRI research as typical text-to-speech algorithms struggle correctly pronouncing the numerous names that exist even just in the English language. This paper presents a pipeline for fusing text and audio features to extract and reuse user information like names with the correct pronunciation. We discuss technical guidelines for implementation and remaining challenges.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Natural language interfaces*; *Sound-based input / output*; • **Computing methodologies** → **Natural language processing**.

KEYWORDS

Name pronunciation, Phoneme Recognition, Named Entity Recognition, Natural Language Processing, Human-Robot Interaction

ACM Reference Format:
James Kennedy, Naveen Kumar, and Maike Paetzel-Prüsmann. 2024. Name Pronunciation Extraction and Reuse in Human-Robot Conversations. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640689

1 INTRODUCTION

Social human-robot interaction has great potential to improve people's lives through applications in healthcare [8], education [4], and entertainment [30], among others. Personalization is a key driver for success in these scenarios; personalized content leads to improved learning gains [5] and improved health outcomes [17], while social personalization, like entrainment, can improve perceptions of robots [15] or build rapport [19].

Calling users by their own (or a chosen) name is a simple technique for personalization that is common in today's chatbot technology [1]. When it comes to social Human-Robot Interaction (HRI), incorporating a user's name has also been suggested as a technique to create a socially engaged character [20]. However, making references to people's names or other personal information like their home town comes with a greater risk of making a social error [28] because modern text-to-speech (TTS) technology still struggles with the correct pronunciation of many named entities. Fig. 1 shows such a use case.

Recent developments in the fields of Natural Language Processing (NLP) and speech processing offer off-the-shelf models for tasks like Named-Entity Recognition (NER) [18], phoneme recognition [32], and speech synthesis [14]. These technologies provide opportunities for designers and technologists to personalize interactions

(a) Robot **asks for but does not use user's name** in dialogue. Conversation feels impersonal, user feels disconnect.

(b) Robot **reuses** the user's name, but **mispronounces** it. User feels disconnected, potentially mocked.

(c) Robot **reuses** the user's name with **correct pronunciation**. Conversation feels personal, user feels heard.

Figure 1: Minimal conversation to demonstrate the effect of name extraction and reuse in social human-robot dialogue.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03
https://doi.org/10.1145/3610978.3640689

with robots and artificial agents, like customizing the phonemes, that is the pronunciation, the TTS applies to individual words.

In this paper, we show how a robot can reuse commonly available features from its audio input streams to extract the information necessary for its TTS to correctly reuse the pronunciation of personal information like names or places. We describe how all the necessary algorithms can be packaged into a single module that can operate in conversational real-time and that can be plugged into common robot control frameworks like the Robot Operating System (ROS) as well as into custom pipelines. Our focus is on providing flexibility to support numerous signals of interest, or different approaches to extracting the relevant parts of the speech and give detailed guidance on the implementation options. We conclude the paper by discussing remaining challenges for extracting and reusing pronunciation information from user speech for personal information in general and names in particular.

2 RELATED WORK

Many aspects of personalization and adaptation have been studied in human-robot interaction, e.g., [15, 19, 22]. Generally, these studies have identified positive effects in the outcomes of interaction task goals or perceptions of the robots. In this work, we focus on meta-information within a conversation with a robot or artificial agent. Rather than personalizing the content, we are interested in modifying the way the agent speaks to improve the naturalness and perception by users.

There are many techniques by which the meta-information in an interaction might be personalized by a robot. One example is through entrainment, the phenomenon by which utterances of speakers in a conversation become increasingly similar to each other. Between people, higher levels of entrainment have been shown to improve mutual likeability [23] and perceived competence [27]. However, it is unclear how much control people have over this entrainment process [7]. Within HRI scenarios, researchers have found that humans will entrain to robots in the lexical content they use [12], and that people will also exhibit synchrony in movement [11]. These phenomena can be explicitly controlled for and encouraged in the case of artificial agents to enhance personalization.

In [13], the authors identify people with the goal of personalizing interactions, but the names given to users are predetermined. In this work, we seek to learn the name directly from a person, and pronounce it correctly. We present this in the context of a pipeline that can be applied to fusing audio and textual features for the purposes of robot reuse, which could find many personalization applications in HRI conversations, e.g., through speech entrainment.

A fundamental building block for using content like names for such personalization is Named Entity Recognition (NER). This enables information extraction from text input, both identifying the type and text associated with different information units, such as names, dates, or locations [21]. Recent progress in neural networks and text embeddings has led to great improvement in the performance of NER in many languages [36], resulting in several robust off-the-shelf models being available that have high accuracy in extracting information like people's names.

Spoken entity extraction adds another layer of complexity to NER due to issues arising from natural speaking styles and errors in speech recognition. As a result there have been recent attempts at joint optimization for end-to-end (E2E) spoken entity extraction using seq2seq architectures. These models often jointly optimize for accuracy of entity extraction [3, 9, 29, 35]. For instance, the authors in [25, 26] develop an E2E model towards spoken name capture, where users say their name and spell it out. One of the main challenges for their model is to be robust to Automatic Speech Recognition (ASR) errors, given the challenges of fine-tuning ASR models specifically for such tasks.

3 PROBLEM STATEMENT - CORRECT PRONUNCIATION OF USER INFORMATION

Names present a prominent case of users providing personal information in a social dialogue with the robot having the opportunity to reuse this information and personalize the interaction to signal understanding. However, a great variety of names exist, causing difficulties for ASR systems to reliably transcribe them, especially in cases with minimal context (like hearing just the name without framing language like "My name is [name]"). Even with a correct transcription, many names can be homographs – the word(s) are spelled the same but have different pronunciations – which means that when a robot attempts to repeat the name, ambiguity remains.

An example conversation is shown in Fig. 1. The robot first asks the user for their name, they respond, then the robot either does or does not reply with their name. Not using the name can be seen as insufficient social skills, or be attributed to a misunderstanding [28]. Reusing the name immediately makes the conversation feel more personal. However, the name "Caroline" has at least two possible pronunciations, represented as /ˈkæ.rə.ˈlaɪn/ or /ˈkæ.rə.ˈlɪn/ using the International Phonetic Alphabet (IPA), i.e., whether the sound at the end is 'lin' or 'line'. The ASR transcript that would typically be used in a dialogue system does not encode this information, so when the string would be used in response, the default TTS pronunciation would be applied, which may be incorrect (Fig. 1 (b)). While likely a good approximation, to truly personalize, having this be fully accurate is desirable for the user to not feel some level of disconnect and potentially even being mocked by the robot.

While the user's name is a common and obvious use case for ensuring a robot uses the same pronunciation as offered by the user, this can be extended to other information like where the user is from, or references to artifacts from their personal life, like toys or places with specific names.

4 IMPLEMENTATION

A schematic of our implementation can be seen in Fig. 2. At a high-level, the audio input, i.e., a user speech utterance, is processed through both speech-to-text and an NLP component to identify the named-entity of interest, and an audio feature extractor to receive the phonemes. The results are then further processed and re-aligned before being passed for storage and inclusion in an output generation step. In the following, we introduce the individual components in more detail.

4.1 Component Overview

Audio Input. For extracting and using a name, the input is assumed to contain a single channel of audio with an utterance from

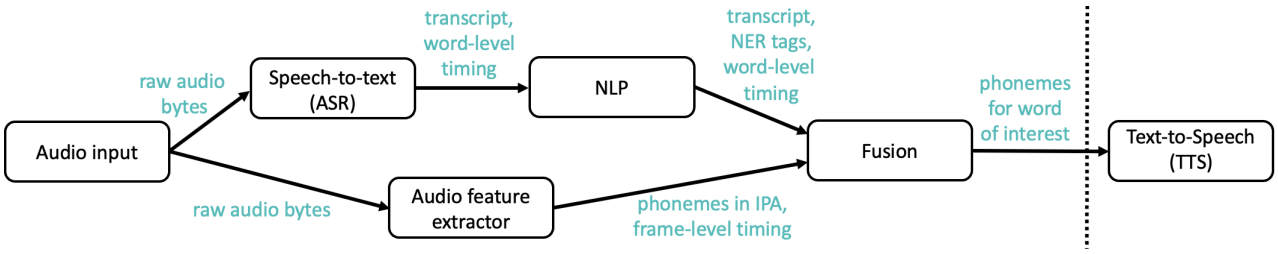

Figure 2: Processing diagram of the pipeline proposed in this paper for extracting and reusing information captured from user speech. The text-to-speech component is not part of our pipeline and only demonstrates the reuse within dialogue.

a user that includes a reference to an entity we would like to store the correct pronunciation for. Going forward, as an example, we will assume this entity to be the user's name. The incoming audio is fed to both an ASR component and to an audio feature extractor.

Speech-to-text (ASR). In our approach, models accepting continuous streams are run for both phoneme recognition and ASR. We require the ASR result to contain word-level timestamps, which is a commonly available feature in off-the-shelf models. The ASR transcription with the aligned timing information is then fed to a language processing component.

NLP. Within the NLP component, a Named-Entity Recognition Flair model is running [24], but any NER could be used. The ASR transcription is passed through the NER to identify names. For each name detected, the string of the token found, its token index (i.e., its position within the utterance), and the character start/end indices will be returned. This then needs realigning to the ASR result, as the ASR result contains the timing from the audio stream. In practice, the tokenization used by ASR and NER are often different. In the example given in Fig. 1, the ASR would return the user speech with the following indices:

Hi,	I'm	Caroline
0	1	2

The NER, however, tokenizes the text as follows:

Hi	,	I'm	Caroline
0	1	2	3

Hence, in this example, the NER will return the index of the recognized name, Caroline, as 3, which is an index that does not exist in the original ASR tokenization. A loop over the ASR words helps to find the matching string and index (in this case 2) and then extract correct word-level timestamps based on this index. The text string along with the word-level timestamps is handed off to the fusion component.

Audio feature extractor. In this case, the audio feature extractor is running a wav2vec2 phoneme prediction model [33]. We use a multilingual phoneme recognition model trained on the Common-Voice dataset [34] to be robust to name pronunciations from different backgrounds. The wav2vec2 model produces phoneme predictions in IPA for every 20ms window of audio. Given a time interval in the audio corresponding to a named-entity, CTC decoding [10] can be applied to predict the phonemes P of the spoken form. Consider the utterance "My name is Caroline" shown in Fig. 2. The audio signal S corresponding to the utterance undergoes feature extraction $X = \mathcal{F}(S)$ which leads to a feature sequence $\{X_1, X_2, \ldots, X_n\}$. After forwarding these through the phoneme recognizer we obtain phoneme posteriors $\{Y_1, Y_2, \ldots, Y_n\}$ for each 20ms window shift through the audio signal. This information along with the word-level timestamps is handed off to the fusion component.

Fusion. From the ASR result, the start and stop timestamp in the audio stream for the word of interest can be determined. This is then fused with the results from the phoneme stream to extract phonemes corresponding to the spoken entity. Specifically, we convert word-level timestamps (T_s, T_e) for the named-entity to indices and perform a CTC decoding over the posterior sequence contained in the interval as follows:

$$s = \lfloor T_s/20 \rfloor; e = \lfloor T_e/20 \rfloor$$
$$P = \{P_1, P_2, \ldots, P_k\} = CTC(Y_s, \ldots, Y_e)$$

The phonetic pronunciation can now be stored, and later injected into a synthesis call to a text-to-speech engine.

Text-to-speech (TTS). The TTS will render audio for playback that uses the name with the pronunciation correctly extracted from the user. Many commercial providers of text-to-speech systems, e.g., Amazon Polly [1], Microsoft Azure [2], and Google [3], allow for custom phonemes to be provided. While the TTS component itself is not part of our pipeline, we are including it in this description for completeness and to showcase how this information is applied in a dialogue.

4.2 Incorporation into Dialogue Framework

The previous section described the general components that are part of our proposed pipeline that receives audio as an input and provides a word of interest with its correct pronunciation as an output. This section will go into more detail specific to the implementation that (a) enables true alignment of the data streams, and (b) minimizes latency when packaging this pipeline into a module within a larger dialogue framework. These are needed for both accuracy, and for a more natural robot behavior.

When fusing any streaming data, time-stamping and alignment is key for accuracy. This becomes challenging if the system is formed

[1] https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html#phoneme-tag
[2] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-synthesis-markup-pronunciation
[3] https://cloud.google.com/text-to-speech/docs/ssml#phoneme

of multiple components, or nodes, running on different computing hardware. For robot dialogue systems that are developed using Microsoft's Platform for Situated Intelligence (psi) [6], the alignment of input streams from different components is already a built-in-feature that can be used directly. While the Robot Operating System (ROS), the framework most commonly used for operating (social) robots, offers some synchronization through the message filter package [31], when working with messages covering vastly different scales (like ASR which spans multiple seconds, and phoneme extraction which covers 20ms windows), this will break down. To solve this, we propose attaching an ID to every audio packet that is streamed, and include this ID (or set of IDs) in every message that generates a result from that audio. It is then straightforward to fuse messages downstream based on the original audio that they were created from, even if they arrive asynchronously. The edge case for this is if the audio packet size is smaller than any downstream processing, or if the results span partial audio packets. The easiest solution is to carefully select the packet sizes to reduce these issues.

To minimize latency, we recommend doing as much processing prior to the fusion as possible. A sequential pipeline approach would be to buffer the audio as it is sent to the ASR, then on receipt of the ASR result, identify the name with NER, use the ASR result to get the timestamps, and send the relevant buffered audio for phone extraction. However, this sequential processing incurs the most latency. Instead, we propose continually running the phoneme extraction in parallel to the ASR calls. As the phoneme extraction is faster to return than ASR, the results will be available as soon as the ASR result arrives. This parallelizes the processing and reduces latency by about 50ms (using NVIDIA RTX 2080), at the cost of increased compute requirements (as phoneme extraction is continually running). In our testing (using Mac M1), NER and alignment takes approximately 15ms, which is therefore the total additional latency added by our name extraction pipeline.

Our pipeline is not required to run in either ROS or psi; it can be integrated into any custom framework. If components like an ASR exist already, then reusing its output by subscribing to the messages the existing ASR already produces minimizes the processing that is required within our pipeline. As our proposed pipeline impacts only a robot's verbal interaction, it can be applied to both physical and virtual embodiments with no adjustments necessary.

5 DISCUSSION

The previous section introduced a pipeline for extracting information of interest from an incoming audio stream and saving the pronunciation of this information for future use. We described how this applies to the user name 'extract and repeat' use-case. Our pipeline has been implemented in an internal prototype into a multi-user HRI setting. We were able to successfully collect and reuse user names during the conversation in order to address the right person. The following discussion summarizes the remaining challenges we experienced with this use-case.

While the solution provided in the previous section for named entity extraction and repetition in general, and for names in particular, will work well in many cases, there are still some challenges that are not addressed. The first of these involves speech impediments. If a user has a speech impediment like a stutter or a lisp, we would not want to reproduce that in the robot speech as it could be perceived as mocking. While ASR results will typically remove phenomena like repetitions, the phoneme extraction of audio would still contain this information. If this pronunciation is reused in the TTS later on, it would repeat the speech impediment. Simple heuristic approaches like removing repeated phones would likely result in other names failing. Instead, more advanced techniques, like speech disfluency prediction [16], may need to be incorporated into the fusion to produce robust behavior. Similarly, if a user is speaking with a different accent from the robot, reusing the mismatching accent in the pronunciation could be perceived as mockery or as culturally insensitive. In this case, the fusion component would require additional processing and could do a heuristic substitution to match from the user's input accent to the character's target accent.

Secondly, if names are not currently detected by the NER model, the approach will fail. This happens more often when minimal context is provided, i.e., the user replies with just their name and no framing language like "My name is". However, it can also happen with names not found in the English NER.

Thirdly, it is possible that multiple pronunciations of the same term occur in one ASR result. For example, a user could be correcting the robot by saying "It's not Carolein, it's Caroline" using the wrong pronunciation first in mimicry of the robot before offering the correct one. In this case, the NLP component needs to be extended to detect not only where an NER occurs, but also which occasion matches the correct phonemes (in this case, the second).

Finally, the NER is reliant on the ASR transcript. This means that the name still needs to be transcribed with enough accuracy for it to be recognized as a name. For instance, the German name 'Maike' is successfully recognized as a name by the NER we use, but is regularly mistranscribed by ASR. This can also produce an effect where if the mistranscription is for another name, e.g., 'Micah', the extraction would succeed with the correct pronunciation, but associated with the wrong written form. Having these models work in tandem can be a challenge, so doing the NER in a single step is increasingly being explored in research [2]. Running multiple ASR or NER components trained or fine-tuned on different languages could increase the accuracy of transcriptions and recognition of names, especially if the number of target languages is small.

6 CONCLUSION

This paper presented an approach for a robot, or other conversational agent, to extract name pronunciation information from a user's utterance and use it in a spoken response. The implementation is described in detail so that others in the field can reproduce the techniques deployed. We discuss the remaining challenges with generalizing to users speaking with different accents or using words from different languages, and for users with speech impediments. The implementation we proposed in this paper can easily be integrated into all common and most custom dialogue frameworks and offers an easy way of integrating more personalization into dialogue in social HRI.

ACKNOWLEDGMENTS

We thank Samantha Ho for the artwork in Fig. 1.

REFERENCES

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. An overview of chatbot technology. In *IFIP international conference on artificial intelligence applications and innovations*. Springer, 373–383.

[2] Guillaume Baril, Patrick Cardinal, and Alessandro Lameiras Koerich. 2022. Named Entity Recognition for Audio De-Identification. In *2022 International Joint Conference on Neural Networks (IJCNN)*. 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892285

[3] Frédéric Béchet, Allen L Gorin, Jeremy H Wright, and Dilek Hakkani Tür. 2004. Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How May I Help You? sm, tm. *Speech Communication* 42, 2 (2004), 207–225.

[4] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science robotics* 3, 21 (2018), eaat5954.

[5] Benjamin S Bloom. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher* 13, 6 (1984), 4–16.

[6] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Patrick Sweeney, Anne Loomis Thompson, and Eric Horvitz. 2021. Platform for Situated Intelligence. arXiv:2103.15975 [cs.AI]

[7] Tanya L Chartrand and John A Bargh. 1999. The chameleon effect: The perception–behavior link and social interaction. *Journal of personality and social psychology* 76, 6 (1999), 893.

[8] Carlos A Cifuentes, Maria J Pinto, Nathalia Céspedes, and Marcela Múnera. 2020. Social robots in therapy and care. *Current Robotics Reports* 1 (2020), 59–74.

[9] Sahar Ghannay, Antoine Caubrière, Yannick Estève, Nathalie Camelin, Edwin Simonnet, Antoine Laurent, and Emmanuel Morin. 2018. End-to-end named entity and semantic concept extraction from speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 692–699.

[10] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*. 369–376.

[11] Syed Khursheed Hasnain, Ghiles Mostafaoui, Robin Salesse, Ludovic Marin, and Philippe Gaussier. 2013. Intuitive human robot interaction based on unintentional synchrony: A psycho-experimental study. In *2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*. IEEE, 1–7.

[12] Takamasa Iio, Masahiro Shiomi, Kazuhiko Shinozawa, Takahiro Miyashita, Takaaki Akimoto, and Norihiro Hagita. 2009. Lexical entrainment in human-robot interaction: Can robots entrain human vocabulary?. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3727–3734.

[13] Bahar Irfan, Michael Garcia Ortiz, Natalia Lyubova, and Tony Belpaeme. 2021. Multi-Modal Open World User Identification. *Transactions on Human-Robot Interaction* 11, 1 (2021). https://doi.org/10.1145/3477963

[14] Navdeep Kaur and Parminder Singh. 2023. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review* 56, 7 (2023), 5837–5880.

[15] Jacqueline M Kory-Westlund and Cynthia Breazeal. 2019. Exploring the effects of a social robot's speech entrainment and backstory on young children's emotion, rapport, relationship, and learning. *Frontiers in Robotics and AI* 6 (2019), 54.

[16] Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad. 2020. Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6089–6093.

[17] Min Kyung Lee, Junsung Kim, Jodi Forlizzi, and Sara Kiesler. 2015. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 743–754.

[18] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* 34, 1 (2020), 50–70.

[19] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2021. Effects of adapting to user pitch on rapport perception, behavior, and state with a social robotic learning companion. *User Modeling and User-Adapted Interaction* 31 (2021), 35–73.

[20] Hadi Beik Mohammadi, Nikoletta Xirakia, Fares Abawi, Irina Barykina, Krishnan Chandran, Gitanjali Nair, Cuong Nguyen, Daniel Speck, Tayfun Alpay, Sascha Griffiths, et al. 2019. Designing a personality-driven robot for a human-robot interaction scenario. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 4317–4324.

[21] David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30, 1 (2007), 3–26.

[22] Joe Saunders, Dag Sverre Syrdal, Kheng Lee Koay, Nathan Burke, and Kerstin Dautenhahn. 2015. "Teach me–show me" end-user personalization of a smart home and companion robot. *IEEE Transactions on Human-Machine Systems* 46, 1 (2015), 27–40.

[23] Antje Schweitzer and Natalie Lewandowski. 2014. Social factors in convergence of F1 and F2 in spontaneous speech. In *Proceedings of the 10th international seminar on speech production, cologne*. 391–394.

[24] Stefan Schweter and Alan Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. arXiv:2011.06993 [cs.CL]

[25] Karan Singla, Shahab Jalalvand, Yeon-Jun Kim, Ryan Price, Daniel Pressel, and Srinivas Bangalore. 2022. Seq-2-Seq based Refinement of ASR Output for Spoken Name Capture. *arXiv preprint arXiv:2203.15833* (2022).

[26] Karan Singla, Yeon-Jun Kim, Ryan Price, Shahab Jalalvand, and Srinivas Bangalore. 2023. E2E Spoken Entity Extraction for Virtual Agents. *CoRR* (2023).

[27] Richard L Street Jr. 1984. Speech convergence and speech evaluation in fact-finding interviews. *Human Communication Research* 11, 2 (1984), 139–169.

[28] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 2 (2021), 1–32.

[29] Natalia Tomashenko, Antoine Caubrière, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2019. Recent advances in end-to-end spoken language understanding. In *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Ljubljana, Slovenia, October 14–16, 2019, Proceedings 7*. Springer, 44–55.

[30] John Vilk and Naomi T Fitter. 2020. Comedians in cafes getting data: evaluating timing and adaptivity in real-world robot comedy performance. In *Proceedings of the 2020 ACM/IEEE international conference on human-Robot Interaction*. 223–231.

[31] ROS Wiki. 2018. message_filters wiki.ros.org. https://wiki.ros.org/message_filters#Time_Synchronizer. [Accessed 21-11-2023].

[32] Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. *CoRR* abs/2109.11680 (2021). arXiv:2109.11680 https://arxiv.org/abs/2109.11680

[33] Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition. *arXiv preprint arXiv:2109.11680* (2021).

[34] Qiantong Xu, Alexei Baevski, and Michael Auli. 2023. Wav2Vec 2.0 LV-60 ESpeak CV Fine-Tuned Model. https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft. Accessed on: 2023.

[35] Hemant Yadav, Sreyan Ghosh, Yi Yu, and Rajiv Ratn Shah. 2020. End-to-end named entity recognition from english speech. *arXiv preprint arXiv:2005.11184* (2020).

[36] Vikas Yadav and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470* (2019).