Interspeech 2024
1-5 September 2024, Kos, Greece

Measuring acoustic dissimilarity of hierarchical markers in task-oriented dialogue with MFCC-based dynamic time warping

Natalia Morozova[1,3], *Guanghao You*[1,3], *Sabine Stoll*[*1,3], *Adrian Bangerter*[*2]

[1]Department of Comparative Language Science, University of Zurich, Switzerland
[2]Institute of Work and Organizational Psychology (IPTO), University of Neuchâtel, Switzerland
[3]Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich

natalia.morozova@uzh.ch

Abstract

Joint activities (e.g. building a LEGO model) unfold in a hierarchy of subprojects. Navigating them implies horizontally elaborating on a subproject (placing one block) and vertically moving to a new subproject (next block). Interactants coordinate horizontal and vertical transitions with project markers (*okay, yeah*). We suggest that vertical vs. horizontal transitions are distinguished both lexically and acoustically. We predicted that acoustic features of identical markers used for different transitions ($okay_{vertical}$ vs. $okay_{horizontal}$) would exhibit more dissimilarity than markers used for same transitions ($okay_{vertical}$ vs. $okay_{vertical}$). We used MFCC-based dynamic time warping to measure dissimilarity between vocalisations and analysed them with a Bayesian regression model. We find that Vietnamese speakers use both lexical and acoustic cues to mark transitions, and paired $same_{horizontal}$ markers are acoustically more similar than $same_{vertical}$ and different-transition markers.

Index Terms: feature extraction, feature matching, mal-frequency coefficients, dynamic time warping, task-oriented dialogue

1. Introduction

Cooperation lies at the heart of human interaction [1], and language constitutes a system of conventions to efficiently navigate cooperative, or joint, activities [2]. Joint activities consist of hierarchies of projects and subprojects [2]. For example, interactants might decide to prepare a three-course meal together. This joint project consists of the subparts, or subprojects, such as preparing course one, two, and three. The subproject of each course might in turn consist of the subsubprojects of individual preparation steps, e.g. washing, chopping, mixing. Interactants need to coordinate how to move from one part of the project to the next. [3] proposed that there are two kinds of generic transitions in the joint action hierarchy. **Horizontal** transitions involve continuing to the next step *within* a project or subproject (e.g., moving from washing ingredients to chopping). **Vertical** transitions involve moving *between* two subprojects by ending a subproject or starting one. This hierarchy also applies to conversations where participants need to coordinate the change of topics and digressions [2].

Interactants typically mark horizontal and vertical transitions implicitly, via words like *uh-huh*, *yeah*, *right*, *all right*, or *okay*. These markers have been previously studied as back-channels [4], continuers [5], acknowledgment tokens [6], or discourse markers [7]. [3] proposed to regroup these words into the category of **project markers**, suggesting that different lexical forms may be specialised for marking **vertical** vs. **horizontal** transitions. Specifically, they found that *okay* and *all right* are vertical transition markers in English and German, whereas *uh-huh*, *mm-hm*, *yeah*, or *right* are horizontal markers. However, these specialisations are probabilistic rather than exclusive, i.e. *okay* is more likely to mark vertical transitions, yet it may also be found in the vicinity of horizontal transitions. Here, we suggest that interactants might use the acoustic modality to further disambiguate the difference between vertical and horizontal transition contexts.

In this paper, we investigate project marker use in Vietnamese task-oriented dialogues. We analyse (1) which lexical forms of project markers interactants use to mark vertical and horizontal transitions and (2) whether they use distinct acoustic features to discriminate between vertical and horizontal functions of identical lexical forms. We predict that (1) while some lexical forms are specialised for marking one transition type, all forms would appear in both horizontal and vertical transitions; and that (2) paired instances of the same lexical marker used in **different** transitions ($okay_{\text{vertical}}$ vs. $okay_{\text{horizontal}}$) have more distinct acoustic features than in **same** transitions ($okay_{\text{vertical}}$ vs. $okay_{\text{vertical}}$ and $okay_{\text{horizontal}}$ vs. $okay_{\text{horizontal}}$). We manually annotated project markers and the transitions where they occurred (vertical vs. horizontal). To compare the acoustic features, we extracted mel-frequency cepstral coefficients (MFCCs) and used dynamic time warping (DTW) alignment as a length-robust method of measuring dissimilarity between two vocalisations. Lastly, we used a Bayesian mixed-effects linear regression model to compare the MFCC-based distances between pairs of markers used in $\mathbf{same_{vertical}}$ and $\mathbf{same_{horizontal}}$ transitions with pairs of markers used in **different** transitions.

The current study suggests that marking of vertical vs. horizontal transitions might occur in both lexical and acoustic modalities. Such findings might have application in automatic speech processing of conversational data, where speech act recognition still remains a challenge. Further, this study advances our understanding of the transition-specific use of project markers for languages beyond English and German [3].

2. Related work

Short words like *okay*, *alright*, *yeah*, and *uh-huh* fulfill a variety of functions in a dialogue. A solid body of research on the function-specific properties of these words indicates that speakers use a combination of lexical and prosodic cues to disambiguate their conversational role.

2.1. Function-specific lexical forms

While each project marker might be used for a variety of pragmatic contexts, English speakers show function-specific biases

* These authors share last authorship.

10.21437/Interspeech.2024-2204

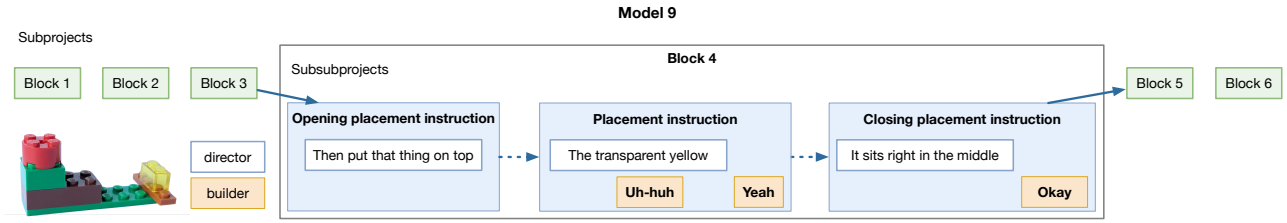

Figure 1: *Hierarchical structure of model-building task with vertical (solid-line arrows) and horizontal (dotted-line arrows) transitions.*

in their lexical choices. Listeners frequently use *okay* and *yeah* as **tokens of incipient speakership** [6, 8], i.e. to acknowledge previous utterances while informing the speaker about an attempt of floor transfer. In contrast, *uh-huh* is used by listeners as non-disruptive **backchannel** to indicate attention and encourage the current speaker to carry on [6, 9]. *Yeah* is the most frequent **yes-answer** and **assessment token** used to express (dis)agreement, acceptance, or rejection of a statement [9]. Among these various project markers, *okay* and *all right* are specialised to **vertically** mark a change of projects, while *yeah* and *uh-huh* **horizontally** mark the continuation within projects [3]. The tendency to use *okay* as a vertical marker has also been documented in other languages [10, 11], where *okay* indicates a change of conversational topics or closes digression sequences and conversations. Despite the associations between lexical forms and pragmatic contexts, many project markers occur in both horizontal and vertical transitions (albeit with different frequencies).

2.2. Function-specific prosodic features

Prosody may help to disambiguate the communicative intention behind project markers. Backchannel *yeah* is shorter in duration and lower in intensity than assessment and agreement *yeah* [9, 12, 13]. Opening and closing *okay* [14, 15] has a flat tone and falling pitch contour, short in duration, and lower in intensity. In contrast, *okay* that encourages the current speaker to continue or requests more information has a rising pitch slope, higher intensity, and longer duration. Some studies reported form-invariant prosodic features associated with specific conversational contexts. For example, English backchannels are described as having higher pitch, intensity, longer duration than other response tokens [16]. Further, response tokens used at turn-changing points generally have a flat-rising tone, as shown for *okay* and *uh-huh* [17]. While these observations suggest that acoustic features of project markers in English may serve to coordinate interactions, to our knowledge no studies have analysed the acoustic properties of these words within the project marker framework (vertical vs. horizontal transitions).

Prosody has been widely used in automatic speech recognition (ASR) pipelines to improve the performance of neural network models. The majority of studies relied on the prosodic features of preceding speech to predict listener responses [18, 19, 20, 21], yet few papers used the prosody of response tokens themselves to disambiguate their contextual functions in ASR [12, 22, 23, 24]. [25] showed that training classifiers on the spectral features of speech with mel-frequency cepstral coefficients (MFCCs) rather than prosody could also lead to high classification accuracy. MFCCs-based recognition of response tokens was successfully implemented for non-lexical reactive tokens [22] and markers of acknowledgement [23]. While the highest classification accuracy was achieved with a combination of acoustic, lexical, and/or contextual features, both studies [22, 23] showed that MFCCs of vocalisations might convey cues about their conversational meaning.

In line with previous findings, we propose that Vietnamese speakers use a combination of lexical and acoustic cues to distinguish between vertical and horizontal transition markers. In this paper, we test specifically if MFCCs of lexically identical markers capture the difference between **same**- and **different**-transition markers in task-oriented dialogues.

3. Data and methods

3.1. LEGO building dataset

To ensure the comparability of our results, we used one of experimental tasks analysed in [3]. We adapted the LEGO building paradigm from [26][1] and collected the recordings from 80 native Vietnamese speakers (40 males, M_{age} = 19.5, SD =1.36) arranged in 20 same-sex and 20 different-sex dyads. Sessions were filmed on university campus and supervised by a Vietnamese assistant. Participants cooperatively built 10 LEGO models in dyads of directors and builders. Directors instructed builders how to construct models based on prototypes that only directors could see. Dyads built 5 models in each of two conditions: visible (no barrier) and hidden (separated with a barrier). All prototype models consisted of six LEGO blocks arranged into abstract shapes. Participants were recorded with video cameras and two lavalier microphones attached to their shirts.

3.2. Data annotation

Following [3], we separated the task routine into a hierarchy of subprojects and subsubprojects (Figure 1). Each model was treated as a project with six nested subprojects for each model block (Block 1, Block 2, etc.). Within each subproject, participants went through subsubprojects of placement instructions. Transitions *between* phases, subphases, and block subprojects were treated as **vertical**. Transitions *within* subsubprojects of placement instructions were labelled as **horizontal**. Vertical transition sequences between block subprojects started with the **closing** placement instruction for one block (e.g. Block 3) and ended with the **opening** placement instruction for the next block (Block 4).

We manually transcribed and annotated data in ELAN [27] (Table 1). Within *Subphase* level, we separated regular, successive identification and placement of blocks (**PL**) from pre-construction block identifications (**ID**) and post-construction model checks (**CH**). We excluded **ID** and **CH** subphases from further analysis due to the uncertainty in their hierarchical or-

[1]The research project proposal was approved by the ethics committee.

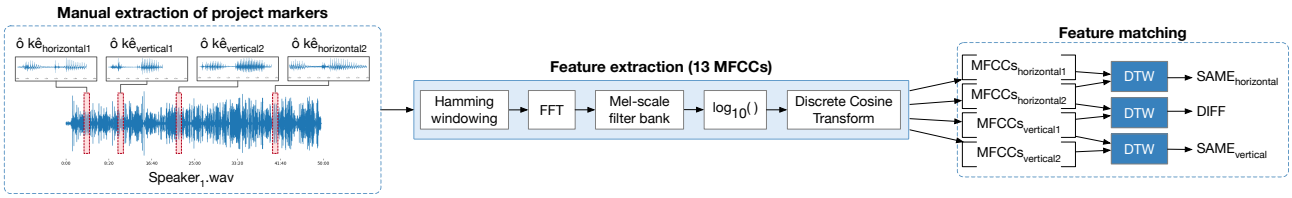

Figure 2: *Mel-frequency cepstral coefficients (MFCCs) extraction and pair-wise Dynamic Time Warping (DTW) pipeline.*

ganisation. In total, we annotated 9037 project markers and their transition-specific functions (vertical or horizontal).

Table 1: *Annotation scheme and inter-rater reliability scores.*

IRR (κ)	Tier	Labels
0.99	Condition	visible, hidden
0.99	Model	m1, ..., m10
0.9	Phase	Entry, Main body, Exit
0.93	Subphase	ID, PL, CH
0.9	Block	b1, ..., b6
0.85	Q&A	q, a
0.82	Transition	vertical, horizontal

3.3. Mel-frequency cepstral coefficients (MFCCs) and Dynamic Time Warping (DTW)

For acoustic analysis, we used 20 sessions of high-quality audio recordings (with 20 male and 20 female participants) and segmented their project marker vocalisations. Few segments containing considerable background noise (e.g. car honking, bell ringing) or laughter were manually filtered out (286 out of 3816) as noise could compromise the accuracy of MFCCs [28]. The final sample included 3530 vocalisations of the highest audio quality and comprised the most frequent marker forms (*rồi*, *ô kê* , *ừm*, *ừm-hứm*, *ờ*, *ừ*, *đúng rồi*). Next, we extracted 13 MFCCs per vocalisation, with sampling rate 16000 Hz, window size 20ms, and 10ms shift using the librosa [29] library.

Since the duration of response tokens is a function-discriminative feature [13, 14, 16], we wanted to preserve the original duration of vocalisations. To do so, we used dynamic time warping (DTW) as a length-robust feature-matching method. DTW computes the minimal alignment path between time series of different lengths by stretching or shrinking the time axis [30], which makes it a suitable method for natural speech processing architectures. Out of 3530 vocalisations, we extracted 66109 normalised DTW distances with dtw-python [1.3.1] library [31]. The distances were computed in a pair-wise manner, between all combinations of identical marker forms (within-marker) produced by the same speakers (within-speaker; Figure 2). Distances between same-transition markers ($okay_{\text{vertical}}$ vs. $okay_{\text{vertical}}$ and $okay_{\text{horizontal}}$ vs. $okay_{\text{horizontal}}$) were labelled as **same$_{\text{vertical}}$** and **same$_{\text{horizontal}}$** and between different-transition markers ($okay_{\text{vertical}}$ vs. $okay_{\text{horizontal}}$) as **diff**.

3.4. Model fitting

We predicted that the distances between identical lexical forms used for different transitions (**diff**) would be greater than between the identical forms used for same transitions (**same$_{\text{vertical}}$** and **same$_{\text{horizontal}}$**). We fitted a Bayesian mixed-effects linear regression in R 4.3.1 [32] with the brms [v2. 20.4] package [33], specifying DTW-based distances as response variable, pair-wise configuration (**diff**, **same$_{\text{vertical}}$**, and **same$_{\text{horizontal}}$**) as fixed-effect predictor, and participant and marker as random effects. We used default model priors.

$$NormalisedDistance(d) \sim skew - Normal(\mu_i)$$
$$\mu_i = \beta_{PairedTransitions} + \alpha_{participant} + \alpha_{marker} \quad (1)$$

The maximum tree depth was set to 15 and delta was adapted to 0.95. Posterior check plots showed good convergence (Pareto $\kappa < 0.4$), and acceptable convergence statistics ($\hat{R} \leq 1.01$).

4. Results

4.1. Lexical forms of Vietnamese project markers

We identified ten categories of lexical forms and their relative frequencies in our dataset: *rồi* (‘*all right*’, 27.1%), *ô kê* (‘okay’, 20.8%), *ừm / ừm-hứm* (‘*mm*, *uh-huh*’ 12.7%), *ờ / ừ* (‘*yeah*’, 12.4%), repetitions of previous utterances or summary statements (12.1%), *đúng / đúng rồi* (‘*right*’, 9.5%), *được / được rồi* (‘*right*, *okay*’, 3%), *vâng*, *dạ*, or *dạ vâng* (‘*yes*’, 1.3%), *chuẩn* (‘*exactly*’, 0.6%), and *chính xác* (‘*correct*’, 0.5%). Some of the least frequent forms are respectful, formal response tokens used to address elders or strangers [34], and thus were rarely used in our task.

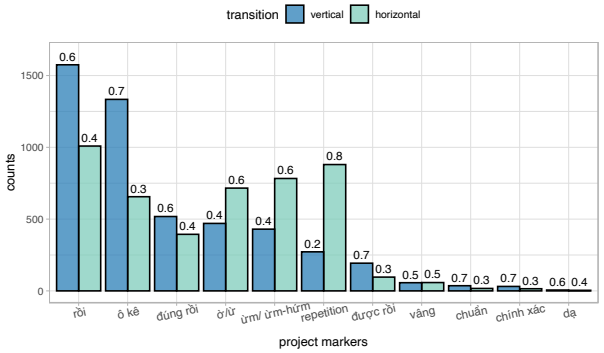

Figure 3: *Total frequency counts of Vietnamese project markers and their ratio distributions between horizontal and vertical transition contexts.*

Consistent with [3], many project markers are specialised for particular transition contexts (Figure 3); e.g. *ô kê* and *được rồi* appear more frequently in vertical transitions, whereas *repetitions* mostly occurred in the proximity of horizontal transitions. Nevertheless, all those project marker forms were used to navigate both transition types, which suggests that the choice of a lexical form alone might not always be a reliable marker of the type of transition. We thus explored the prediction that identical lexical forms might display more distinct acoustic features

when they mark different transition points (**diff**) as opposed to same transitions (**same**$_{\textbf{vertical}}$ and **same**$_{\textbf{horizontal}}$)

4.2. Acoustic distances between paired vocalisations

Figure 4 summarises the central tendencies in posterior pair-wise distance distributions and differences in group means between **diff**, **same**$_{\textbf{vertical}}$, and **same**$_{\textbf{horizontal}}$ pair-wise configurations. The model output indicates that the bulk of alignment distances lies within the interval bounds of 50 and 75 for all pair-wise configurations.

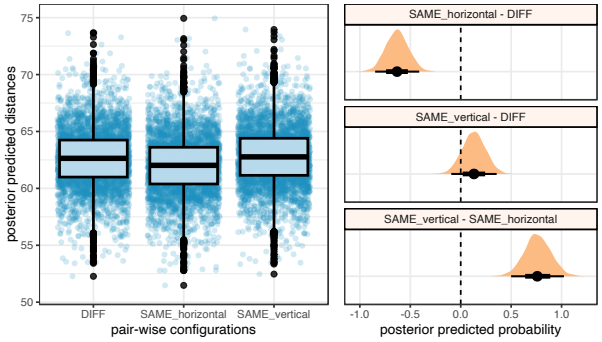

Figure 4: *Posterior predictions of pair-wise distance distributions and contrasts between their group means.*

The spectral features of paired forms of **same**$_{\textbf{horizontal}}$ vocalisations showed less acoustic variability, and thus lower mean distances, than different-transition markers (**diff**) (HDI 95% = [-0.85, -0.41]). The model also indicated the positive difference between the mean distances of paired **same**$_{\textbf{vertical}}$ and **same**$_{\textbf{horizontal}}$ forms, estimating greater variability in acoustic features of vertical transition markers (HDI 95% = [0.5, 1.02]).

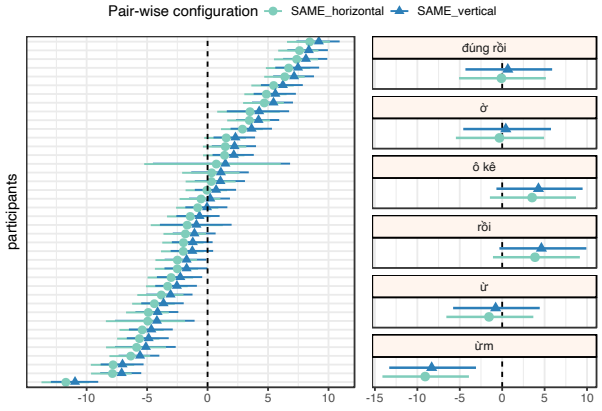

Figure 5: *By-participant and by-marker slope variation relative to the intercept (**diff**).*

Figure 5 shows between- and within-participant variation relative to the intercept (**diff**). For the majority of participants, paired lexical forms for **different** transitions were pronounced with more distinct acoustic features than for **same** transition contexts. Additionally, by-marker variation suggests that acoustic features of project markers might depend on their lexical forms. Markers *ừm* and *ừ*, which are top frequent horizontal markers in our dataset (Figure 3), appear to be more acoustically similar in **same**-transition contexts, yet we observe the opposite trend for predominantly vertical markers *ô kê* and *rồi*.

5. Discussion

We examined the hypothesis that interlocutors use a combination of lexical and acoustic cues to navigate two generic transitions: vertical and horizontal. We have described the lexical forms of Vietnamese project markers and their distribution across two transition contexts. Consistent with the findings of [3], our results show that some lexical forms of project markers are predominantly used for certain transition contexts (e.g. vertical *ô kê* and horizontal *repetitions*). However, ultimately every lexical form was used for both transitions.

To test for differences in the acoustic modality, we used dynamic time warping (DTW) to measure dissimilarity between mel-frequency cepstral coefficients (MFCCs) of Vietnamese project markers. We found that paired instances of horizontal markers (**same**$_{\textbf{horizontal}}$) are produced with less variation in spectral features, suggesting that interlocutors acoustically distinguish horizontal markers with a specific set of features. Paired marker forms used for vertical transitions (**same**$_{\textbf{vertical}}$), on the other hand, exhibit greater acoustic dissimilarities. As vertical transitions mark both *openings* and *closings* of subprojects, i.e. two inherently distinct progress states, it is possible that interlocutors additionally distinguish *entries* into new subprojects from *exits*.

Lastly, not all lexical forms show equal variation in their acoustic markings, suggesting an interplay between transition-specific biases in lexical and acoustic cues. It might be that lexical forms which frequently appear in a variety of contextual functions (e.g. signalling **continuations**, **openings**, and **closings** of subprojects like *rồi*) generally exhibit greater acoustic variation than the markers that are primarily specialised for a specific interaction context (e.g. just to mark **continuations**). In follow-up studies, we aim to explore which particular lexical and prosodic features distinguish vertical markers from horizontal.

A limitation of this study, however, is that we focused only on spectral acoustic features. While the distances between MFCCs alignment paths should indicate the changes in frequency bands of two vocalisations that could be detected by the human ear, they tell us little about the particular acoustic features influencing those differences such as pitch and its contour, energy, jitter, etc. Overall, the results of our study suggest that project markers are widely used by interlocutors to support the structure of their interaction and the function of these markers can be disambiguated by their acoustic features.

6. Acknowledgements

The authors would like to express their gratitude to the University of Languages and International Studies (Vietnam National University, Hanoi) for granting us the opportunity to conduct data collection within their facilities and for their support in data annotation. Furthermore, we thank anonymous reviewers for their constructive feedback. This work was supported by the NCCR Evolving Language, Swiss National Science Foundation Agreement no. 180888.

7. References

[1] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[2] H. H. Clark, *Using Language*, ser. "Using" Linguistic Books. Cambridge University Press, 1996.

[3] A. Bangerter and H. H. Clark, "Navigating joint projects with dialogue," *Cognitive Science*, vol. 27, no. 2, pp. 195–225, 2003.

[4] V. H. Yngve, "On getting a word in edgewise," in *CLS-70*. University of Chicago, 1970, pp. 567–577.

[5] E. A. Schegloff, "Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences," in *Analyzing Discourse: Text and Talk*, D. Tannen, Ed. Washington, D.C.: Georgetown University Press, 1982, pp. 71–93.

[6] G. Jefferson, "Notes on a systematic deployment of the acknowledgement tokens "yeah" and "mm hm"," *Papers in Linguistics*, vol. 17, pp. 197–206, 1984.

[7] D. Schiffrin, *Discourse markers*. Cambridge: Cambridge University Press, 1987.

[8] K. Drummond and R. Hopper, "Back channels revisited: Acknowledgment tokens and speakership incipiency," *Research on Language and Social Interaction*, vol. 26, no. 2, pp. 157–177, 1993.

[9] D. Jurafsky, E. Shriberg, B. Fox, and T. Curl, "Lexical, prosodic, and syntactic cues for dialog acts," in *Discourse Relations and Discourse Markers*, 1998, pp. 114–120.

[10] E. Betz and A. Deppermann, "Okay in responding and claiming understanding," in *OKAY across languages: toward a comparative approach to its use in talk-in-interaction*, ser. Studies in language and social interaction, E. Betz, A. Deppermann, L. Mondada, and M.-L. Sorjonen, Eds. Amsterdam: John Benjamins Publishing Company, 2021, vol. 34, pp. 56–92.

[11] L. Mondada and M.-L. Sorjonen, "Okay in closings and transitions," in *OKAY across languages: toward a comparative approach to its use in talk-in-interaction*, ser. Studies in language and social interaction, E. Betz, A. Deppermann, L. Mondada, and M.-L. Sorjonen, Eds. Amsterdam: John Benjamins Publishing Company, 2021, vol. 34, pp. 94–127.

[12] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, and C. van Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[13] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *Proc. Interspeech 2010*, 2010, pp. 3058–3061.

[14] W. A. Beach, "Using prosodically marked "okays" to display epistemic stances and incongruous actions," *Journal of Pragmatics*, vol. 169, pp. 151–164, 2020.

[15] E. Couper-Kuhlen, "The prosody and phonetics of okay in american english," in *OKAY across languages: toward a comparative approach to its use in talk-in-interaction*, ser. Studies in language and social interaction, E. Betz, A. Deppermann, L. Mondada, and M.-L. Sorjonen, Eds. Amsterdam: John Benjamins Publishing Company, 2021, vol. 34, pp. 132–173.

[16] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in american english," in *Proceedings of the 16th International Congress of Phonetic Sciences*, 2007, pp. 1065–1068.

[17] B. A. Hockey, "Prosody and the role of okay and uh-huh in discourse," in *Proceedings of the Eastern States Conference on Linguistics*, 1993, pp. 128–136.

[18] A. I. Adiba, T. Homma, and T. Miyoshi, "Towards immediate backchannel generation using attention-based early prediction model," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7408–7412.

[19] A. Gravano, S. Benus, H. Chávez, J. Hirschberg, and L. Wilcox, "On the role of context and prosody in the interpretation of 'okay'," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, A. Zaenen and A. van den Bosch, Eds., 2007, pp. 800–807.

[20] R. Ruede, M. Müller, S. Stüker, and A. Waibel, "Enhancing backchannel prediction using word embeddings," in *Proc. Interspeech 2017*, 2017, pp. 879–883.

[21] K. P. Truong, R. Poppe, I. de Kok, and D. Heylen, "A multimodal analysis of vocal and visual backchannels in spontaneous dialogs," in *Proc. Interspeech 2011*, 2011, pp. 2973–2976.

[22] T. Kawahara, K. Sumi, Z.-Q. Chang, and K. Takanashi, "Detection of hot spots in poster conversations based on reactive tokens of audience," in *Proc. Interspeech 2010*, 2010, pp. 3042–3045.

[23] D. Neiberg and K. P. Truong, "Online detection of vocal listener responses with maximum latency constraints," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5836–5839.

[24] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–374, 2000.

[25] D. Ortega, C.-Y. Li, and N. T. Vu, "Oh, jeez! or uh-huh? a listener-aware backchannel predictor on asr transcriptions," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8064–8068.

[26] H. H. Clark and M. A. Krych, "Speaking while monitoring addressees for understanding," *Journal of Memory and Language*, vol. 50, no. 1, pp. 62–81, 2004.

[27] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds., 2006, pp. 1556–1559.

[28] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[29] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, K. Huff and J. Bergstra, Eds., 2015, pp. 18 – 24.

[30] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[31] T. Giorgino, "Computing and visualizing dynamic time warping alignments in r: The dtw package," *Journal of Statistical Software*, vol. 31, no. 7, pp. 1–24, 2009.

[32] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[33] P.-C. Bürkner, "brms: An r package for bayesian multilevel models using stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[34] T. D. Ha, X. V. Ha, N. N. Nguyen, T. T. Le, and D. Starks, "Yeah in vietnamese english: A study of l2 interviewer speech," *The European Journal of Applied Linguistics and TEFL*, vol. 7, no. 1, pp. 91–106, 2018.