The 34th Conference on Computational Linguistics and Speech Processing (ROCLING 2022)
Taipei, Taiwan, November 21-22, 2022. The Association for Computational Linguistics and Chinese Language Processing

Analyzing discourse functions with acoustic features and phone embeddings: non-lexical items in Taiwan Mandarin

Pin-Er Chen
National Taiwan University
cckk2913@gmail.com

Yu-Hsiang Tseng
National Taiwan University
seantyh@gmail.com

Chi-Wei Wang
National Taiwan University
r09142007@ntu.edu.tw

Fang-Chi Yeh
National Tsing Hua University
fangchiyeh2000@gmail.com

Shu-Kai Hsieh
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

Non-lexical items are expressive devices used in conversations that are not words but are nevertheless meaningful. These items play crucial roles, such as signaling turn-taking or marking stances in interactions. However, as the non-lexical items do not stably correspond to written or phonological forms, past studies tend to focus on studying their acoustic properties, such as pitches and durations. In this paper, we investigate the discourse functions of non-lexical items through their acoustic properties and the phone embeddings extracted from a deep learning model. Firstly, we create a non-lexical item dataset based on the interpellation video clips from Taiwan's Legislative Yuan. Then, we manually identify the non-lexical items and their discourse functions in the videos. Next, we analyze the acoustic properties of those items through statistical modeling and building classifiers based on phone embeddings extracted from a phone recognition model. We show that (1) the discourse functions have significant effects on the acoustic features; and (2) the classifiers built on phone embeddings perform better than the ones on conventional acoustic properties. These results suggest that phone embeddings may reflect the phonetic variations crucial in differentiating the discourse functions of non-lexical items.

Keywords: non-lexical item, discourse function, acoustic property, acoustic representation, pragmatics

1 Introduction

People's everyday interactions include sounds that are not verbal words in the traditional sense. These sounds, such as sighs, sniffs, and grunts, are used in indexing the turn-taking in dialogues, marking stance, showing affections, and expressing roles and meanings in conversations (Dingemanse, 2020). Examples of these *non-lexical items* are *un-huh* in English as a marker showing understanding and attentiveness, while the single syllable *uh* and *um* act as fillers and disfluency markers (Ward, 2006; Buschmeier et al., 2011).

While these non-lexical items are important linguistically, they pose an interesting challenge to linguistic inquiry. Non-lexical items do not belong to a major word class, and some do not conform to the language's phonological requirements (Keevallik and Ogden, 2020). Moreover, while the phonetic properties of non-lexical items could be generally described, they are nevertheless "phonetically underspecified."(Keating, 1988) For example, in the study of "moan" in board game interactions, Hofstetter (2020) found "moans" involve phonetic properties related to open vowels, irrespective of their frontness, backness, or roundedness. The study suggests that a non-lexical item can not be represented as a single phonetic symbol; instead, it may refer to the vowel space for which we do not have a general phonetic symbol. Some studies, therefore, analyze these items in terms of their acoustic properties: the components' sound (Ward, 2006), the fundamental frequencies, durations, and intensities. (Shan, 2021; Ballier and Chlébowski, 2021).

In contrast to the conventional acoustic property analysis, an alternative approach to analyzing non-lexical items is through the acoustic representations learned by data-

driven methods. These methods include deep learning models mapping the audio segments to the latent embedding space from acoustic data in a (self-)supervised fashion (Li et al., 2020; Xu et al., 2021; Baevski et al., 2020). Although the models are not explicitly trained to represent the similarities among phonetic features, studies nonetheless find the audio segments with similar linguistic properties are closer together in the embedding space (Ma et al., 2021; Cormac English et al., 2022; Silfverberg et al., 2021). Therefore, these phonetic representations may already encode the phonetic variability of non-lexical items to reflect their different discourse functions.

This study thus aims to investigate how the acoustic properties contribute to the non-lexical items' discourse functions and how the phone embeddings extracted from the deep learning model help differentiate those functions. The rest of the paper is organized as follows. We first review related works on discourse markers and how they are analyzed with acoustic properties (Sec. 2). Next, we describe our dataset on non-lexical items (Sec. 3) in Taiwan Mandarin, in which we manually identify the items and annotate their discourse functions in interpellation video clips of Taiwan's Legislative Yuan. Finally, based on the dataset, we conduct the acoustic property analysis (Sec. 4) and build classifiers based on the phone embeddings extracted from a deep learning model (Sec. 5). Finally, Section 6 concludes the paper.

2 Related Works

2.1 Discourse Marker

Discourse markers (hereafter, DMs) has received increasing attention since Schiffrin (1987, p. 31) initially defined them as "sequentially dependent elements which bracket units of talk." However, little consensus has been not only on the terminology[1] of DMs but on the classification frameworks. Schiffrin (1987) has proposed that DMs form a category composed of phrases, conjunctions, and interjections, and that they have a part in discourse coherence considering different planes of talk.[2] Additionally, DMs can also serve as identifiers of participation status, speaker's assumptions, or hearer's knowledgement (Schiffrin, 1987; Schwenter, 1996; Fraser, 1999).

Despite that earlier research considered DMs as text-connective items bonding to syntactic structures, Fischer (2006, p. 9) defined DMs as devices involved in "turn-taking, interpersonal management, topic structure, and participation frameworks." Subsequently, Diewald (2006, 2013) suggested that DMs demonstrate pragmatic functions, manage discourse in a syntactically-independent way, and present their polyfunctionality in discourse (c.f. Fraser, 2009; Hansen, 2006; Németh, 2022).

Although numerous analyses were conducted on the pragmatic functions of DMs, they focused mostly on the associations with semantic senses and syntactic structures (e.g., Aijmer, 2011; Crible, 2017; Ford and Thompson, 1996). That is, studies of the connections between the discourse functions and the phonological information of DMs are relatively few.

2.2 Acoustic Property

The previous works which interwove DMs and their acoustic properties were mainly on the pragmatic-prosodic interface. Shan (2021) and Zhao and Wang (2019) investigated the Mandarin Chinese DMs, 你知道 *ni zhidao* 'you know' and 你不知道 *ni bu zhidao* 'you don't know', respectively. While Shan (2021) analyzed on duration, tempo, intensity, and fundamental frequencies (i.e., pitch, hereinafter F_0), Zhao and Wang (2019) examined the speech tempo, mean F_0 frequencies, and pitch accents of the DMs. In general, they have found correlations between the discourse functions and the acoustic properties. Moreover, Tseng et al. (2006) have suggested that connectors are predictable from speech prosody; most 'redundant prosodic fillers' are duration-triggered and manifested through

[1] For instance, discourse marker (Jucker and Ziv, 1998; Schiffrin, 1987); discourse particles (Aijmer, 2002; Fischer, 2006); pragmatic marker (Brinton, 1996); among others

[2] Schiffrin has suggested the five planes of talk: the Exchange structure (ES), Action structure (AS), Ideational structure (IdS), Participation framework (PF), and Information state (InS). More details can be seen in Schiffrin (2005), Maschler and Schiffrin (2015), and Hamilton et al. (2015).

narrowed F_0 ranges, whereas 'obligatory discourse markers' are syntax-triggered and manifested through widened F_0 ranges and resets.

The acoustic properties and their relevance to the pragmatic functions of DMs have also been analyzed cross-linguistically (e.g., Cabarrão et al., 2018; Raso and Vieira, 2016; Gonen et al., 2015; Beňuš, 2014). Referring to Wu et al. (2021), the phonetic variations of DMs in French are likely to appear in spontaneous speech and undergo phonetic reduction, considering their shorter mean phone duration and a rather centralized vowel space. Additionally, Schubotz et al. (2015) investigates the common English construction *you know* in terms of its duration, which is likely to be affected by the residuals of speech rate.

In addition to acoustic properties, past studies also examined the phonetic representations learned with data-driven methods. For example, Silfverberg et al. (2021) studied phonological alternations of Finnish consonant gradation with vector representations retrieved from RNN models. Other studies also tried to learn dense vector representations purely from text using grapheme-to-phoneme mappings with CBOW and SkipGram models (O'Neill and Carson-Berndsen, 2019). Notably, recent studies found transformer-based speech processing models (Baevski et al., 2020; Hsu et al., 2021), while not explicitly modeling phonetic properties, encoded the phonetic categorization information in the model representations, such as vowels and consonants, or fricatives and stops (Ma et al., 2021; Cormac English et al., 2022).

Tracing back to the former sections, previous literature on DMs mostly concentrated on their status at the semantic-pragmatic interface. The reviewed acoustic-related research, however, focused on those construction-wise DMs, and not to mention that the analyzed acoustic properties were limited to suprasegmental features, such as pitch and duration. In this case, the potential phonetic-pragmatic interrelationship of non-lexical items is yet to be elaborated.

3 Non-lexical Items Dataset

First, we used four interpellation video clips from Taiwan's Legislative Yuan.[3]. Audio tracks were then extracted from the clips, converted into 16 bit `WAV` format, and resampled with 22kHz sampling rates. The overall data comprise separate interpellation of two male and two female legislators, each ranging 6-8 minutes. The equal number of genders was to balance potential gender differences in the utterances.

Secondly, the audio segments of non-lexical items (e.g., *uh, em,* and *ho*) were annotated by three native speakers via Praat 6.2.03 (Boersma and Weenink, 2021). Each non-lexical item acquired two tags, one for functional *Role* and one for pragmatic *Meaning*. Referring to Ward (2006), we defined the six candidates of *Role* as follows:

- `BACKCHANNEL`, which occurs repetitively and shows the agreement of the hearer; it often overlaps the main channel[4] of the utterance.
- `CFT (Clause-final token)`, which occurs in the sentence-final position and ends certain turn of talk.
- `DISFLUENCY`, which refers to the onset or coda of a word that can hardly be recognized due to its discoursal incompleteness.
- `FILLER`, which serves as a connector between two sentences or a sentence-initial particle of the speaker.
- `RESPONSE`, which occurs in the main channel and often indicates a flippant attitude.
- `OTHER`, which represents the non-lexical item not belonging to the above types.

Similarly, we summarized the following eight candidates for *Meaning*. It is noted that certain non-lexical items may carry multiple pragmatic meanings, and that the candidates below are not mutually exclusive. Thus, one non-lexical item is allowed to be annotated with multiple *Meaning* tags.

[3]The clips were downloaded from the Parliament TV website (https://www.parliamentarytv.org.tw/) and encoded as AAC, H.264

[4]see also Heinz (2003), Li et al. (2010), and McNely (2009) among others.

- `authority.` The speaker demonstrates his profession, personal experience, or intention in the speech.
- `control.` The speaker is in control of knowing exactly what to say or do next.
- `concern.` The speaker lacks confidence in his own words or tries to show respect to the audience.
- `thought.` The speaker takes the words (from himself or the other participant) as involving or meriting thought.
- `dissatisfaction.` The speaker is unsatisfied with his own words, the conversation, or the other participant.
- `new information.` The speaker wants to express that he has received new information; the speaker successfully lets the other participant understand the topic of the speech.
- `old ground.` The speaker is expecting to move on to the next topic since he has already acknowledged the current one.
- `neutral.`

In sum, a total of 143 non-lexical items produced by the legislators were manually annotated. We then moved on to extract the acoustic properties for the dataset.

4 Acoustic Property Analysis

With the assumption that the discourse functions may encode phonological variations, we illustrated our data collection and the annotation for non-lexical items in Sec. 3. The following sections (4.1 and 4.2) then present the analyses and results of acoustic properties.

4.1 Property Extraction

For each non-lexical item, we retrieved six conventional acoustic properties: mean pitch, duration, F1, F2, F3, and nasality, via customized Praat scripts (Styler, 2017). As formant frequencies construct the vowel space, F1 is determined by the vowel height, F2 is determined by the vowel backness, and F3 is determined by the vowel roundness.[5]

In terms of nasality, it can be quantified by ***a1-p1*** (for high vowels such as [i, u, y]) or ***a1-p0*** values (for non-high vowels such as [a, o, ə, e]). Since most of the annotated non-lexical items are realized and transcribed with non-high vowels, only the ***a1-p0*** values were considered. While ***a1*** stands for the amplitudes (in *dB*) of F1, ***p0*** stands for the amplitude of the nasal peak below F1 (Chen, 1997; Cho et al., 2017; Chiu and Lu, 2021).

Subsequently, to build up the most comprehensive acoustic properties, the values of F1, F2, F3 frequencies and ***a1-p0*** amplitude for each annotated non-lexical item were measured at 5 different time-points (i.e., the 10%, 30%, 50%, 70%, 90% time-points within each item interval). The retrieved acoustic data for 715 tokens[6] were processed and modified into machine-readable forms using the `pandas` package (The Pandas Development Team, 2020) in Python 3.8.9 (Python Core Team, 2021).

The statistical analysis was performed via the `lmerTest` package (Kuznetsova et al., 2017) in R 4.2.1 (R Core Team, 2022). Some factors contain rare categories were therefore re-coded. Specifically in the candidates of *Role*, `DISFLUENCY` and `RESPONSE` in were merged into `OTHER`, considering their extremely few occurrences. As for the candidates of *Meaning*, the items with multiple candidate tags were recoded as `complex`. The `OTHER` and `complex` were set as references in *Role* and *Meaning* factors, respectively. Finally, Box-Cox transformations (Box and Cox, 1964) were applied to each response variable to reduce the non-normalities in the distributions.

4.2 Evaluations

To explore the effect of discourse functions on the acoustic properties, we conduct statistical analyses with linear mixed-effects models and classification tasks with SVM.

Statistical Modeling. Apart from the two discourse functions (*Role* and *Meaning*), we also take *Transcriptions* into consideration. As *Transcriptions*, annotated for segment-identification, reflects the annotators' perception for each non-lexical item, it is likely a

[5]The higher the F1, the lower the vowel; the higher the F2, the more anterior the vowel; the lower the F3, the rounder the vowel (Flanagan, 1955; Lindblom and Studdert-Kennedy, 1967).

[6]Each 143 annotated non-lexical items were measured at 5 different points, resulting in 715 tokens.

	Chiq	Df	*p*-value	
Duration	83.79	9	<.001	***
Pitch	124.66	9	<.001	***
F1	10.12	9	.341	
F2	20.32	9	.016	*
F3	7.62	9	.573	
Nasality	15.29	9	.083	

Table 1: Model comparisons of linear mixed-effects in different response variables. The comparisons are between the base model, which only contains transcription and random intercepts, and the full model, which additionally includes discourse function predictors. For brevity, only comparison statistics are shown. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

control variable that poses significant effects on the properties. Thus, for the evaluation of each acoustic property, we actually compare two models: one full linear mixed-effects model (composed of *Role*, *Meaning*, and *Transcriptions*) as well as one counterpart baseline model (composed of only *Transcriptions*).

Table 1 illustrates the sequential (Type I) ANOVA results for the linear mixed-effects models, in which one specific acoustic property is used as the dependent variable. Specifically, the acoustic properties that reach statistical significance among the model comparisons are `Duration`, `Pitch`, and `F2`, suggesting that certain types of roles and meanings present additional effects on acoustic properties, after controlled for the transcriptions. These results imply acoustic properties help differentiate discourse functions.

To further examine such possibility, Table 2 compiles the fixed-effect results of the full linear mixed-effects models for the acoustic properties, where the discourse functions[7] are the predictors. We find that `Pitch` shows the most significance when predicting both discourse functions, which corresponds to the previous works introduced in Sec. 2.2. Yet, `Duration` and `F2` are only capable of predicting certain types of *Meaning* and without any overlap. Not to mention the other three acoustic properties (i.e., `F1`, `F3`, and `Nasality`) which did not show any statistical significance.

To sum up, the overall effectiveness of the linear mixed-effects models for the acoustic properties to predict the discourse functions remain questionable. In the following section, we go on to the implementation of the alternative model, the Support Vector Machines (SVM).

Support Vector Machines Support Vector Machines (SVM) model is implemented for the classification tasks, in which the acoustic properties are used in prediction of discourse functions. As we assume that the discourse functions may reflect in the phonological variations of the non-lexical items, linear models such as SVM are applicable.

We use random 70-30 splits for training and testing data. While the training data comprise 500 tokens, the testing data comprise 215 tokens. A random guessing model, serving as a *the-most-frequent baseline*, is also implemented for comparison. It calculates the frequency distributions of all discourse functions, and then it invariably predicts the most frequent class. We use the accuracy, precision, recall, and F1-score to evaluate the performance of the two models.

Table 3 shows that both models, based on the acoustic properties, find it harder to predict *Meaning* than *Role*. Specifically, the `acoustics` achieved slightly better accuracy (.48) and precision (.09) than the baseline (.38 and .04). In the prediction of *Role*, however, the performance of the models was very similar. It implies that the `acoustics` in fact does not acquire much advantage in predicting discourse functions. This observation is consistent with the results of the previous liner mixed-effects model, in which we found few correlations between the acoustic properties and the discourse functions. Therefore, we attempt to find other presentations of phonological variations that may better capture the candidates of discourse functions with higher accuracy.

5 Phone embeddings

As the conventional acoustic properties did not show promising results of capturing the

[7]Notice that the aforementioned `BACKCHANNEL` (as *Role*) and `concern` (as *Meaning*) only exist in the supplementary annotation for those non-lexical items produced by the administrative officers in opposition to the legislators. Data are reserved for the future studies.

	Duration	Pitch	F1	F2	F3	Nasality
(transcriptions)			--			
CFT	0.034	12.04***	35.68	6.28	10 169.4	4.03
FILLER	0.042	14.92***	2.67	1.22	10 913.4	5.67
authority	−0.016	3.87**	3.98	2.29***	−6832.3	2.52
control	−0.013	0.16	3.49	7.87	2345.1	0.18
dissatisfaction	−0.052	−10.07***	45.70	3.16**	−9942.1	4.08
neutral	−0.016	0.05	58.17	1.58*	1948.2	0.30
new information	−0.267**	10.17***	−40.21	1.65	−5134.1	−2.71
old ground	−0.003	0.82	−4.51	1.31	3383.3	0.13
thought	−0.288***	−2.36	97.46	1.55	2643.0	2.75

Table 2: Parameter estimates of discourse functions in the linear-mixed effect models. The variables of transcriptions are included in all models, but their estimates are not shown in the table for brevity. Response variables are Box-Cox transformed, the parameters are therefore in the transformed scale. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Role	Acc	Pr	Rc	F1
acoustics	.76	.15	.20	.17
acoustics-base	.76	.15	.20	.17
Meaning	Acc	Pr	Rc	F1
acoustics	.48	.09	.14	.11
acoustics-base	.38	.04	.10	.06

Table 3: Evaluation of acoustic models

discourse functions, we reached out to phonetic vector representations, in which the phonological variations of non-lexical items might be encoded.

Instead of the common end-to-end models trained on waveforms and language-specific transcriptions in ASR tasks, we chose the *Allosaurus* model by Li et al. (2020)[8] for retrieving the phone embeddings. Specifically, the Allosaurus is an universal phone recognizer integrating an ASR encoder with an allophone layer, in which language-independent phone distributions are directly recognized and mapped into language-dependent phoneme distributions.

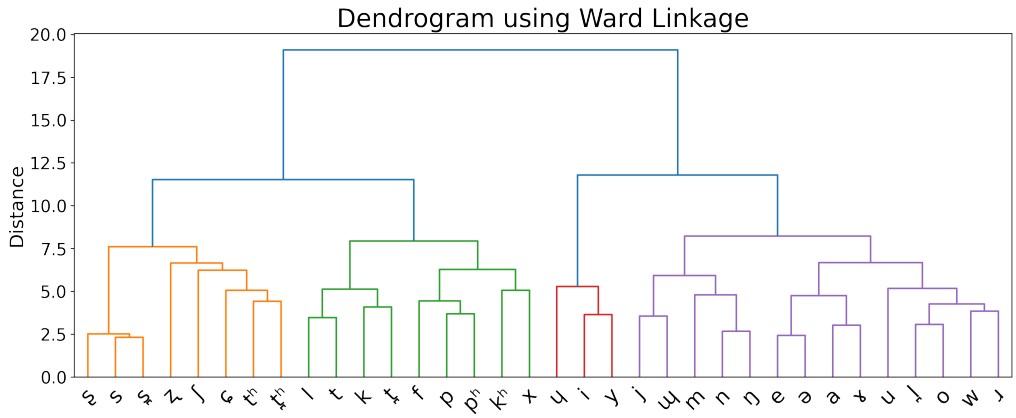
We first examine the phone embeddings learned by the phone recognition model. In the video clips collected in Section 3, the model automatically identifies 29,218 phones in the conversations. To investigate the phone organizations in the embedding space, we then extract the bi-LSTM representations[9] with which model predicts the phones as phone embeddings. Next, we average these embeddings by their predicted phones and obtain 34 phone centroids in the embedding space. We follow the literature (Cormac English et al., 2022) and conduct hierarchical clustering with Ward linkage based on the Euclidean distances between the centroids. The clustering results are shown in Figure 1a and Figure 1b. We not only observe clear clusters of vowels and consonants but observe that the fricatives and stops tend to be close to each other with similar phonetic properties. The patterns suggest that the phone embeddings might reflect the phonetic variations in our conversation data.

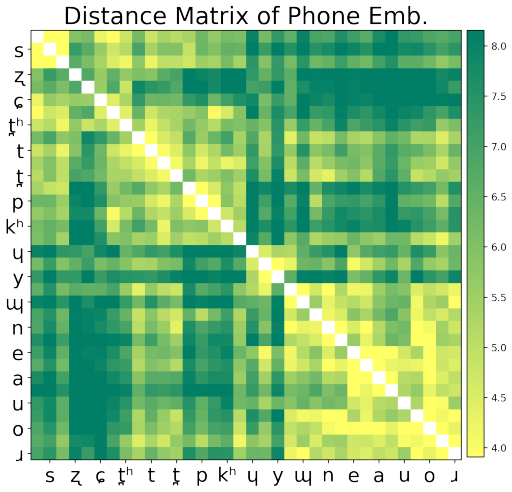
Moreover, we inspect the clustering structure of recognized phones that occurred in the non-lexical items. Figure 1c shows the two-dimensional t-SNE (Pedregosa et al., 2011) visualization of the 640-dimension phone embeddings obtained from Allosaurus. The same phones tend to form distinct clusters, and the general distinction between vowels and consonants is still observed in the figure. It indicates that the embeddings may represent their corresponding phonetic properties. As Li et al. (2020) have shown in their studies, Allosaurus has the advantage of multilingual phone recog-

[8] https://github.com/xinjli/allosaurus

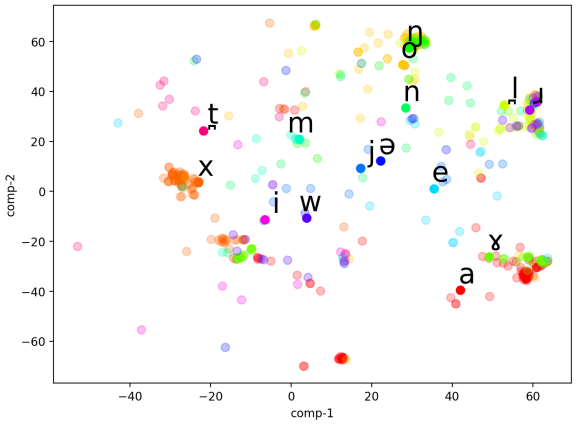
[9] Referring to the comments from the reviewers, the bi-LSTM representations are used as the phone embeddings considering their better performance than the other representations (i.e., the 40-dimension MFCCs and the phone logits) generated by Allosaurus.

(a)





(b)



(c)

Figure 1: (a) The dendrogram of the hierarchical clustering with Ward linkage. The links are color-coded for visual references. Generally, the top left and right branches loosely correspond to consonants and (semi-)vowels. The leftmost branch (orange) are mostly fricatives (e.g., s, ʂ, ɕ); the one on the right (green) includes stops (e.g., k, t, p). (b) The distance matrix shows a consistent pattern with the one in the dendrogram. (c) The t-SNE projection of the phones in non-lexical items. Only the most-frequent 15 phones are shown for clarity. IPA symbols mark the median points of each category.

nition and involves more phonological knowledge. It is thus appropriate for us to leverage these phone embeddings, by which the discourse functions of non-lexical items may be encoded.

5.1 Classification Task

The output data by Allosaurus (i.e., the phone embeddings and phoneme transcriptions) are aligned with our annotations of discourse functions for non-lexical items. It is noted that only the phoneme, whose timestamp matches the 715 tokens of non-lexical items, are kept for the classification tasks. The data is split randomly 70-30 into training and testing datasets as in Section 4.2.

We also implement a linear SVM model and a random guessing model serving as a *the-most-frequent baseline* for the classification tasks.[10] The only difference here is that we replace use the acoustic properties with the phone embedding vectors to predict the candidates of the discourse functions.

[10]Regarding the comments from the reviewers, the

Role	Acc	Pr	Rc	F1
`phone emb.`	.92	.96	.87	.91
`baseline`	.78	.16	.20	.18
Meaning	Acc	Pr	Rc	F1
`phone emb.`	.77	.84	.68	.72
`baseline`	.42	.05	.11	.07

Table 4: Evaluation of classifiers based on phone embeddings

5.2 Evaluation Results

As shown in the upper part of Table 4, `phone emb.` stands out with the highest accuracy (.92) and precision (.96) in prediction of *Role*. While `baseline` presents the accuracy of .78, the acoustic models (see Table 3) show even lower accuracies (.76) and precision (.15). As for predicting *Meaning*, `phone emb.` significantly outperforms its baseline and remains the highest in accuracy (.77) and precision (.84) among all models. In general, `phone emb.` presents superior performance than the other models in prediction of both discourse functions.

Moreover, both models (i.e., `acoustics` and `phone emb.`) are better at predicting *Role* than *Meaning*, likely due to the fact that *Meaning* comprises more types of candidates and internally more equal distribution. In this case, the gap between the accuracies of `phone emb.` (i.e., between .92 and .77) is still the smallest among the models. This suggests that our model is better at capturing the discourse functions by using the phone embeddings, the phonetic realizations, than the statistical acoustic properties.

6 Conclusion

This paper focuses on the phonetic-pragmatic interrelationship of non-lexical discourse markers in Taiwan Mandarin. As we assume that the discourse functions may be captured by the phonological variations, we firstly analyzed on the common acoustic properties (i.e., duration, nasality, mean pitch, F1, F2, and F3), followed by the classification tasks considering the 640d-phone embeddings. In comparison with the conventional acoustic properties, the model using phonetic realizations performs better in prediction of the functional *Role* and pragmatic *Meaning* of the non-lexical items. The result is consistent with our hypotheses that the phonetic realizations, embeddings via deep learning, encode certain phonological variations of non-lexical items and correlate with their discourse functions.

linear SVM model and the model baseline are adopted to not only display the data distributions but highlight the results of Allosaurus, as we mainly focus on whether the phone representations really help us explore non-lexical items. Based on the results, we did find the the model using phonetic realizations performs better in predicting the discourse functions, and we expect future research to develop better representations and state-of-the-art models that allow us to describe non-lexical items more appropriately.

Acknowledgments

We sincerely thank the three native-speakers for the annotation work and the two anonymous reviewers whose comments and suggestions have helped us clarify the technical details in the paper.

References

Karin Aijmer. 2002. *English discourse particles: evidence from a corpus*. Number 10 in Studies in corpus linguistics. Benjamins, Amsterdam.

Karin Aijmer. 2011. Well i'm not sure i think⋯the use of well by non-native speakers. *International Journal of Corpus Linguistics*, 16:231–254.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Nicolas Ballier and Aurélie Chlébowski. 2021. "see what i mean, huh?" evaluating visual inspection of f0 tracking in nasal grunts. In *Interspeech 2021*, pages 376–380. ISCA.

Štefan Beňuš. 2014. Conversational entrainment in the use of discourse markers. In *Recent Advances of Neural Network Models and Applications*, pages 345–352. Springer.

Paul Boersma and David Weenink. 2021. Praat: Doing phonetics by computer [computer program] version 6.2.03, retrieved 23 august 2022 from http://www.praat.org/.

George EP Box and David R Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Laurel J. Brinton. 1996. *Pragmatic markers in English: grammaticalization and discourse functions*. Number 19 in Topics in English linguistics. Mouton de Gruyter, Berlin ; New York.

Hendrik Buschmeier, Zofia Malisz, Marcin Włodarczak, Stefan Kopp, and Petra Wagner. 2011. Are you sure you're paying attention?'-uh-huh'communicating understanding as a marker of attentiveness. In *Twelfth Annual Conference of the International Speech Communication Association*.

Vera Cabarrão, Helena Moniz, Fernando Batista, Jaime Ferreira, Isabel Trancoso, and Ana Isabel Mata. 2018. Cross-domain analysis of discourse markers in european portuguese. *Dialogue & Discourse*, 9(1):79–106.

Marilyn Y Chen. 1997. Acoustic correlates of english and french nasalized vowels. *The Journal of the Acoustical Society of America*, 102(4):2360–2370.

Chenhao Chiu and Yu-An Lu. 2021. Articulatory evidence for the syllable-final nasal merging in taiwan mandarin. *Language and Speech*, 64(4):771–789.

Taehong Cho, Daejin Kim, and Sahyang Kim. 2017. Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in english. *Journal of Phonetics*, 64:71–89.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2022. Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington. Association for Computational Linguistics.

Ludivine Crible. 2017. *Discourse Markers and (Dis) fluency across Registers*. Ph.D. thesis, Université de Berne.

Gabriele Diewald. 2006. Discourse particles and modal particles as grammatical elements. *Approaches to Discourse Particles*, pages 403–425.

Gabriele Diewald. 2013. "same same but different" - modal particles, discourse markers and the art (and purpose) of categorization. *Discourse Markers and Modal Particles.Categorization and Description*, pages 19–46. Cited By :34.

Mark Dingemanse. 2020. Between sound and speech: Liminal signs in interaction. *Research on Language and Social Interaction*, 53(1):188–196.

Kerstin Fischer. 2006. Towards an understanding of the spectrum of approaches to discourse particles: introduction. In Fischer, editor, *Approaches to discourse particles*, number 1 in Studies in pragmatics, pages 1–20. Elsevier, Oxford.

James L Flanagan. 1955. A difference limen for vowel formant frequency. *The journal of the Acoustical Society of America*, 27(3):613–617.

C Ford and S Thompson. 1996. Interactional units in conversation: Syntactic, intonational and pragmatic resources. *Interaction and grammar*, (13):134.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952.

Bruce Fraser. 2009. An account of discourse markers. *International Review of Pragmatics*, 1:293–320.

Einat Gonen, Zohar Livnat, and Noam Amir. 2015. The discourse marker axshav ('now') in spontaneous spoken hebrew: Discursive and prosodic features. *Journal of Pragmatics*, 89:69–84.

Heidi E Hamilton, Deborah Tannen, and Deborah Schiffrin. 2015. *The handbook of discourse analysis*. John Wiley & Sons.

Maj-Britt Mosegaard Hansen. 2006. *A dynamic polysemy approach to the lexical semantics of discourse markers: (with an exemplary analysis of French toujours*, number 1 in Studies in pragmatics, pages 21–41. Elsevier, Netherlands.

Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of pragmatics*, 35(7):1113–1142.

Emily Hofstetter. 2020. Nonlexical "moans" : Response cries in board game interactions. *Research on Language and Social Interaction*, 53(1):42–65.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

A.H. Jucker and Y. Ziv. 1998. *Discourse Markers: Descriptions and Theory*. New series]. Lightning Source Incorporated.

Patricia A. Keating. 1988. Underspecification in phonetics. *Phonology*, 5(2):275–292.

Leelo Keevallik and Richard Ogden. 2020. Sounds on the margins of language at the heart of interaction. *Research on Language and Social Interaction*, 53(1):1–18.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26.

Han Z Li, Yanping Cui, and Zhizhang Wang. 2010. Backchannel responses and enjoyment of the conversation: The more does not necessarily mean the better. *International journal of psychological studies*, 2(1):25.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David Mortensen, Graham Neubig, Alan Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. pages 8249–8253.

Björn EF Lindblom and Michael Studdert-Kennedy. 1967. On the role of formant transitions in vowel recognition. *The Journal of the Acoustical society of America*, 42(4):830–843.

Danni Ma, Neville Ryant, and Mark Liberman. 2021. Probing acoustic representations for phonetic properties. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 311–315.

Yael Maschler and Deborah Schiffrin. 2015. Discourse markers language, meaning, and context. *The handbook of discourse analysis*, pages 189–221.

Brian McNely. 2009. Backchannel persistence and collaborative meaning-making. In *Proceedings of the 27th ACM international conference on Design of communication*, pages 297–304.

Zsuzsanna Németh. 2022. The conversation-organising role of the non-lexical sound öö in hungarian. *Journal of Pragmatics*, 194:23–35.

Emma O'Neill and Julie Carson-Berndsen. 2019. The effect of phoneme distribution on perceptual similarity in english. In *The 20th Annual Conference of the International Speech Communication Association (Interspeech 2019), Graz, Austria, 15-19 September 2019*. ISCA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Python Core Team. 2021. *Python: A dynamic, open source programming language*. Python Software Foundation. Python version 3.8.9.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Tommaso Raso and Marcelo Vieira. 2016. A description of dialogic units/discourse markers in spontaneous speech corpora based on phonetic parameters. *CHIMERA Romance corpora and linguistic studies*, 3:221.

Deborah Schiffrin. 1987. *Discourse Markers*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Deborah Schiffrin. 2005. Discourse markers: Language, meaning, and context. *The handbook of discourse analysis*, pages 54–75.

Louise Schubotz, Nelleke Oostdijk, and Mirjam Ernestus. 2015. Y' know vs. you know: What phonetic reduction can tell us about pragmatic function. In *S. Lestrade, P. de Swart & L. Hogeweg (Eds.). Addenda. Artikelen voor Ad Foolen.*, pages 261–280. Nijmegen: Radboud Universiteit Nijmegen.

Scott A. Schwenter. 1996. Some reflections on o sea: A discourse marker in spanish. *Journal of Pragmatics*, 25(6):855–874.

Yi Shan. 2021. Investigating the interaction between prosody and pragmatics quantitatively: A case study of the chinese discourse marker ni zhidao ("you know"). *Frontiers in psychology*, 12.

Miikka Silfverberg, Francis Tyers, Garrett Nicolai, and Mans Hulden. 2021. Do RNN states encode abstract phonological alternations? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5501–5513, Online. Association for Computational Linguistics.

Will Styler. 2017. On the acoustical features of vowel nasality in english and french. *The Journal of the Acoustical Society of America*, 142(4):2469–2482.

The Pandas Development Team. 2020. *pandas-dev/pandas: Pandas*.

Chiu-yu Tseng, Zhao-yu Su, Chun-Hsiang Chang, and Chia-hung Tai. 2006. Prosodic fillers and discourse markers–discourse prosody and text prediction. In *Tonal Aspects of Languages*.

Nigel Ward. 2006. Non-lexical conversational sounds in American English. *Pragmatics & Cognition*, 14(1):129–182.

Yaru Wu, Mathilde Hutin, Ioana Vasilescu, Lori Lamel, Martine Adda-Decker, Liesbeth Degand, et al. 2021. Fine phonetic details for discourse marker disambiguation: a corpus-based investigation. In *The 10th Workshop on Disfluency in Spontaneous Speech (DiSS 2021)*.

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

Beibei Zhao and Gaowu Wang. 2019. The prosodic features of the mandarin discourse marker nibuzhidao under different functions. In

Proceedings of the 3rd International Conference on Art Design, Language, and Humanities.