Multilingual Turn-taking Prediction Using Voice Activity Projection

Koji Inoue[1], Bing'er Jiang[2], Erik Ekstedt[2], Tatsuya Kawahara[1], Gabriel Skantze[2]
[1]Graduate School of Informatics, Kyoto University, Japan
[2]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
inoue.koji.3x@kyoto-u.ac.jp, binger@kth.se, erikekst@kth.se
kawahara@i.kyoto-u.ac.jp, skantze@kth.se

Abstract

This paper investigates the application of voice activity projection (VAP), a predictive turn-taking model for spoken dialogue, on multilingual data, encompassing English, Mandarin, and Japanese. The VAP model continuously predicts the upcoming voice activities of participants in dyadic dialogue, leveraging a cross-attention Transformer to capture the dynamic interplay between participants. The results show that a monolingual VAP model trained on one language does not make good predictions when applied to other languages. However, a multilingual model, trained on all three languages, demonstrates predictive performance on par with monolingual models across all languages. Further analyses show that the multilingual model has learned to discern the language of the input signal. We also analyze the sensitivity to pitch, a prosodic cue that is thought to be important for turn-taking. Finally, we compare two different audio encoders, contrastive predictive coding (CPC) pre-trained on English, with a recent model based on multilingual wav2vec 2.0 (MMS).

Keywords: Turn-taking, Multilingual, Spoken Dialogue System, Voice Activity Projection

1. Introduction

Turn-taking is a fundamental aspect of spoken interaction between humans, and consequently an important function to model in spoken dialogue systems (Skantze, 2021). In human-human conversations, the transitioning of the conversational floor is smoothly conducted. It has been shown across various languages that the transition offset is typically very brief, around 100-500 msec (Stivers et al., 2009). This indicates that humans use various turn-taking signals across multiple modalities, including lexical cues, prosody, gaze, respiratory, and gestures, in order to coordinate (Włodarczak and Heldner, 2016; Kendrick et al., 2023). In addition, given that the listener also needs some time to articulate a response, there is likely a prediction mechanism involved, where the listener predicts that the speaker's utterance is about to end (Garrod and Pickering, 2015; Levinson and Torreira, 2015; Ishimoto et al., 2017).

While recent advancements in large language models (LLMs) have made it easier to generate highly sophisticated responses in spoken dialogue systems, turn-taking is still typically handled in a very simplistic manner. In practical spoken dialogue systems, turn-taking is commonly implemented using a simple silence timeout threshold, typically around 1 second, to indicate the end of a turn. Silence, however, is not a very good indicator, as silences within turns (pauses) are typically longer than silences between turns, in human-human interaction (Heldner and Edlund, 2010). This means that spoken dialogue systems are often plagued by long response delays or frequent interruptions in pauses.

To address this problem, many proposals have been made for end-of-turn prediction models. These models consider verbal and non-verbal cues (such as linguistic and prosodic features) of preceding user utterances, in order to predict whether the user is just pausing (a *hold*), or whether the turn is yielded (a *shift*). In earlier models, feature engineering was common, but it has now become more popular to input time-series data, such as prosodic features and word vector representations (word embeddings), into neural networks like recurrent neural networks (RNNs) (Skantze, 2017; Masumura et al., 2017). More recently, transformer-based models have been proposed, which can take in the raw input text or audio in an end-to-end processing manner (Ekstedt and Skantze, 2020; Sakuma et al., 2023; Muromachi and Kano, 2023; Kurata et al., 2023).

Another limitation of earlier models has been their sole focus on the binary prediction of turn hold vs. shift. A more comprehensive model of turn-taking should involve more nuanced decisions. When taking the turn, it is, for example, necessary to determine the appropriate waiting time before starting to talk (Raux and Eskenazi, 2012; Lala et al., 2018; Sakuma et al., 2023). There is also a difference in predicting backchannels vs. turn-shifts (Lala et al., 2017). As stated above, humans are able to not just react to turn-yielding cues, but they can also predict upcoming turn-shifts. This would clearly also be a desirable property of spoken dialogue systems. Furthermore, there is no established and robust method for handling interruptions

LREC-COLING 2024, pages 11873–11883

and overlaps in conversation, which are commonly observed in human-human conversations. A more dynamic turn-taking prediction model is required to enable spoken dialogue systems to handle turn-taking in a more human-like manner. Crucial for such models is that they do not make turn-taking decisions at specific events, but that they operate in a continuous fashion.

Several models have been proposed recently that make more nuanced turn-taking predictions continuously in a time frame manner (Skantze, 2017; Lala et al., 2019). Among such continuous models, the voice activity projection (VAP) model is used in this study (Ekstedt and Skantze, 2022b). The VAP model uses multi-layer Transformers and predicts the near future voice activities of dialogue participants by processing the raw audio signals from the two speakers in a dyadic dialogue. Previous work has shown that the VAP model outperforms other models in predicting turn-taking behaviors, including backchannel predictions (Ekstedt and Skantze, 2022b). In the latest version of this model, cross-attention Transformer layers are added after the self-attention layer, to model the audio from the two speakers separately, as explained in Section 2. Recently, the VAP model has been extended for various purposes including backchannel prediction (Liermann et al., 2023), multi-modal turn-taking prediction (Onishi et al., 2023), and its real-time processing (Inoue et al., 2024).

To our knowledge, the VAP model has so far only been trained and tested for English (Ekstedt and Skantze, 2022a). Since it only operates on raw audio, it is technically straightforward to apply it to other languages (even ones it was not trained on). However, it is not clear how much of turn-taking cues are universal. In this paper, we investigate to what extent a model trained on one language can be transferred to other languages, but also whether it is possible to build a multilingual model. This would clearly be more desirable than having separate models trained specifically for different languages. Specifically, we aim to construct a trilingual model for English, Mandarin Chinese, and Japanese. Those languages were partly chosen since they represent three different language families (Germanic, Sino-Tibetan, and Japonic) and therefore should exhibit a certain level of diversity.

Previous research has analyzed the differences in turn-taking behavior among languages. Typical differences include the timing of turn-taking (Stivers et al., 2009; Dingemanse and Liesenfeld, 2022). For example, the turn transition time of Mandarin and Japanese, centering around 0 msec, while English has more overlaps between turns Dingemanse and Liesenfeld (2022). Furthermore, in the analysis of turn-taking cues, it has been pointed out that the intonation change at the end of preceding utterances is effective regardless of the language (Duncan, 1972; Local et al., 1986; Koiso et al., 1998; Ward and Tsukahara, 2000; Levow, 2005; Gravano and Hirschberg, 2011). Specifically, Mandarin shows turn-final pitch lowering for all words in both task-oriented and daily conversations, regardless of the original lexical tone Jian and Wu (2011); Levow (2005). Moreover, there are differences in the use of backchannels (short utterances such as "yeah" and "yes"), which are essential behaviors in turn-taking, and it has been noted that Japanese has the highest frequency, followed by English and then Chinese (Clancy et al., 1996). In summary, there are both common tendencies and differences in turn-taking behavior among languages, which justifies the importance of the proposed multilingual turn-taking prediction model.

Achieving a multilingual turn-taking prediction model can lead to the realization of a multilingual spoken dialogue system that does not require specifying the input language. To our knowledge, this is the first attempt to achieve a multilingual turn-taking model based on the cross-attention Transformer.

Based on the above, this study sets the following research questions:

RQ1: Can a VAP model trained on one language be directly applied to another language?

RQ2: Is it possible to train a single multilingual VAP model that would be on-par with a monolingual model (trained and evaluated on the same language)?

RQ3: Has the multilingual model learned to identify the language?

RQ4: How important is pitch for the multilingual model?

RQ5: What is the effect of the audio encoder on the model's performance?

The rest of this paper is organized as follows: First, in Section 2, we explain the VAP model that serves as the basis for this study. In Section 3, we introduce each of three language dialogue datasets used in this study and analyze the differences among the languages. We conduct experiments to answer the research questions mentioned above in Section 4 and conclude the paper in Section 5.

2. Voice Activity Projection (VAP)

As stated above, the main objective of the VAP model is to predict a future voice activity of two speakers in a dialogue, based on raw audio input. For this study, we use the public repository of the VAP model[1] to make the results reproducible. Note

[1] https://github.com/ErikEkstedt/VoiceActivityProjection

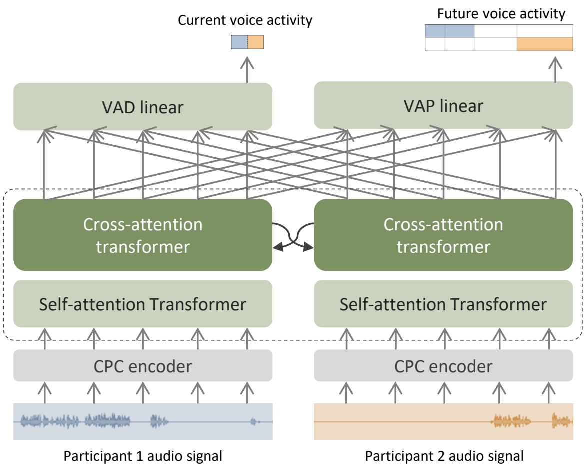


Figure 1: Architecture of the VAP model

that parameters of the VAP model in this study are derived from the above original repository.

2.1. Model Architecture

Figure 1 illustrates the architecture of the VAP model. The input is a stereo audio signal, with each channel corresponding to each participant's audio. The length of the input audio signal is assumed to be a maximum of 20 seconds. In this study, we use a sampling rate of 16 kHz, and a frame rate of 50Hz.

The input signal of each channel is encoded by a pre-trained model of Contrast Predictive Coding (CPC). The CPC model used in this study is composed of a 5-layer CNN and a single-layer GRU, and it is pre-trained with the Librispeech dataset (Riviere et al., 2020). The unsupervised pre-training algorithm of CPC predicts latent representations of audio in the near future, using a contrastive learning, where negative samples are taken from other time frames. In this way, the model is similar to the VAP model, in terms of predicting the near future. Additionally, the pre-trained CPC model has been reported to be useful for multilingual phoneme recognition (Riviere et al., 2020), even when English was used for the pre-training. Thus, we will also rely on a CPC model pre-trained on English. However, in Section 4.4, we also compare the performance with a multilingual audio encoder (MMS). During the training of the VAP model, the parameters of the pre-trained CPC are frozen. The dimension of the vector output by the CPC encoder model is 256.

The vectors encoded by the CPC model are inputted to a self-attention Transformer for each channel. In this study, we utilize a one-layer Transformer with a dimension of 256. Subsequently, the outputs from the two channels are fed into a cross-attention Transformer. In this Transformer, the vector from the first channel serves as the query, while another vector from the second channel acts as the key and value. The reverse case is also simultaneously performed. This way, interactive information between the two channels are encoded. The output is the combined result of these two computations. This cross-attention mechanism draws inspiration from recent dialogue audio generative models like dGSLM (Nguyen et al., 2023). In our current setup, we employ three-layer Transformers for this interactive mechanism, with a dimension of 256. The final output of this Transformer is the concatenation of the dual Transformers, resulting in a dimension of 512. The number of attention heads is set to 4, and the dropout rate during training is set at 0.1.

Finally, the output vector is passed through two separate linear layers for multitask learning. The first main task is the VAP objective itself, with a dimension of 256 (see next section). The second task is voice activity detection (VAD), a subtask that detects the current voice activities of the two participants. The output vector of this subtask has two dimensions, where each dimension corresponds to the voiced probability of each participant. The future voice activity depends on the current voice activity, so by adding this subtask of VAD, we aim to stabilize the training of VAP.

2.2. VAP State

The main objective of the VAP model is to predict the voice activity for both participants within a two-second time window. Instead of making independent predictions for both speakers (as in (Skantze, 2017)), the VAP model makes a prediction of the joint activity of the two speakers over the future time window. The time window is divided into four binary bins: 0-200 msec, 200-600 msec, 600-1200 msec, and 1200-2000 msec, as depicted in Figure 2. Since the model assumes two speakers, there are a total of 8 binary bins. This results in 256 ($= 2^8$) combinations of possible activations, representing various events such as turn-shifts, backchannels, overlapping speech, etc. The objective for the VAP model is formulated as a multi-class classification problem that aims to predict which state out of the 256 patterns the future two seconds will fall into, and the model will output a probability distribution over those states.

Since the ground truth voice activity is often recorded with finer step sizes than those bins, they have to be discretized. Figure 2 illustrates this process, where a bin is defined as "voiced" if there are more voiced frames than unvoiced frames in it, and "unvoiced" otherwise.

2.3. Loss Function

The outputs of the VAP and VAD tasks are applied to the softmax function, and then probabilities of

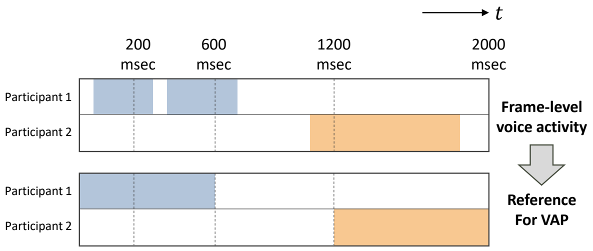


Figure 2: Discretizing bins for the VAP model

VAP state indexed as $y \in (1, \cdots, 256)$ and voice activity of participant s are calculated as $p_{vap}(y)$ and $p_{vad}(s)$, respectively. Then, the cross-entropy losses with respect to the reference data are computed as:

$$L_{vap} = -\log p_{vap}(y) \ , \tag{1}$$

$$L_{vad} = -\sum_{s=1}^{2} \{v_s \log p_{vad}(s) + (1 - v_s)\log(1 - p_{vad}(s))\} \tag{2}$$

where y is the index of the reference VAP state, and $v_s \in (0, 1)$ is the reference voice activity of participant s (1 for voiced, 0 for unvoiced).Finally, the losses of both VAP and VAD are combined by adding them together to form the final loss function for optimization as

$$L = L_{vap} + L_{vad} \ . \tag{3}$$

Note that the notation for time frames is omitted due to space limit, although the calculations mentioned above are performed for all input time frames.

2.4. Turn-taking Prediction Using VAP

While the probability distribution over the possible VAP states represents a complex prediction of what the turn-taking dynamics will look like in the near future, it can be hard to use and interpret directly. A simplified representation of the output can be obtained by summing up the probability values of each participant's bins in the 0-200 msec and 200-600 msec regions. Then, softmax can be applied to both sums to obtain $p_{now}(s)$, which represents a short-term future voice activity prediction of each participant (i.e., "how likely is each participant to speak in the next 600 msec"). Similarly, for the 600-1200 msec and 1200-2000 msec bins, $p_{future}(s)$ is used as a slightly longer-term future voice activity prediction. It is important to note that this is just one example of how the VAP output can be utilized.

3. Datasets

In this study, we use three dyadic conversational datasets: English, Mandarin, and Japanese, to investigate a model for multilingual turn-taking.

3.1. Switchboard (English)

The Switchboard dataset is a collection of telephone conversations recorded in English, covering everyday topics (Godfrey et al., 1992). It consists of a total of 2,438 dialogues, equivalent to approximately 259.1 hours of data. This dataset was divided into training, validation, and test sets in a ratio of 8:1:1 at the session level, using random selection. The training set comprises 1950 dialogues (218.7 hours). The validation and test sets contain 244 dialogues (20.2 hours), respectively. Since not all datasets were of equal size, and in order to make the comparison between languages fair, we selected a subset of the data to be used. Therefore, the training and validation sets were further randomly sub-sampled to approximately 92.5 and 11.5 hours, in order to align them to the size of the smallest dataset, namely the Japanese dataset.

3.2. HKUST Mandarin Telephone Speech

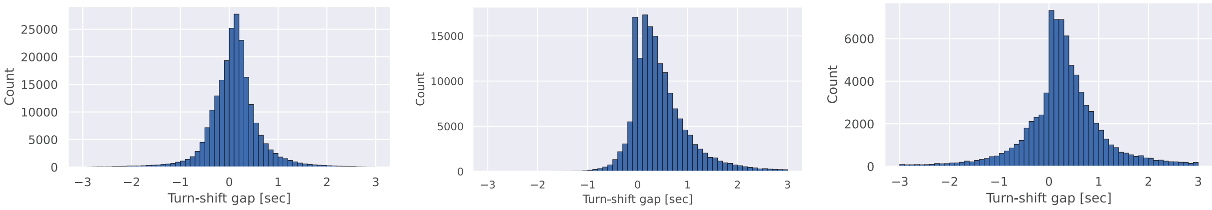
The HKUST Mandarin telephone speech corpus is a collection of Mandarin Chinese spoken dialogues in telephone conversations (Liu et al., 2006), similar to Switchboard. The dataset contains a total of 867 dialogues, approximately 148.6 hours in duration. These have been divided into the training of 758 dialogues (approximately 130.1 hours), the validation of 88 dialogues (14.5 hours), and the test of 24 dialogues (3.9 hours), respectively. Similar to Switchboard, in order to match the size of the smallest dataset, which is the Japanese dataset, the training and validation sets have been sub-sampled to 92.5 hours and 11.5 hours, respectively.

3.3. Travel Agency Task Dialogues (Japanese)

Travel Agency Task Dialogues is a project that collects simulated dialogue data for travel consultations in Japanese (Inaba et al., 2022). These conversations were recorded using an online conference system. The dialogues simulate online conversations between a travel agency staff and a customer, with the role of the staff being played by someone with actual experience working in a travel agency. Note that while the participants were assigned roles, they were not given a script. A total of 329 dialogues (115.5 hours) have been recorded. These dialogues were randomly divided into session units, with 263 dialogues (92.5 hours) in the training set, and 33 dialogues (11.5 hours) each in the validation set and test set.

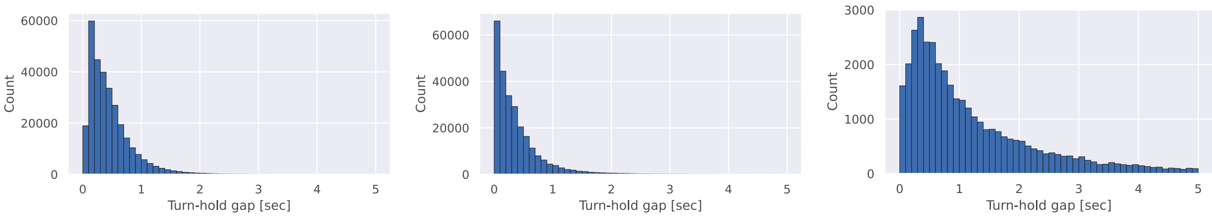
3.4. Differences Across Languages

As mentioned above, earlier studies suggest that there are slight differences in turn-taking tendencies



(a) English (Switchboard) (b) Mandarin (HKUST) (c) Japanese (Travel agency)

Figure 3: Histogram of **turn-shift** gap in three languages



(a) English (Switchboard) (b) Mandarin (HKUST) (c) Japanese (Travel agency)

Figure 4: Histogram of **turn-hold** gap in three languages

between the three languages. Therefore, we first investigated the differences in gap and pause length between the datasets. Figure 3 shows the histogram of gap duration during turn transitions, while Figure 4 illustrates the distribution of pause length during turn holding. It can be observed that Mandarin and Japanese tend to have slightly shorter gaps during turn transitions compared to English. On the other hand, during turn holding, English and Mandarin show a tendency towards shorter pause lengths, compared to Japanese. Furthermore, it's worth observing that the distribution of Japanese is more evenly dispersed. The Japanese dataset exhibits a formal dialogue setting with explicitly assigned roles, distinguishing it from the other two datasets. Consequently, there seems to be a trend in the Japanese dataset where the participants hold the turn for longer periods of time without turn-shifts, resulting in longer gaps between utterances.

Given these differences, it is important to note that the VAP model does not only have to take into account how cues and signals may differ between languages, but also the overall distributions, which might bias the predictions.

4. Experiments

To answer the research questions from **RQ1** to **RQ5** mentioned in Section 1, we conducted a series of experiments, as described below.

4.1. Cross-lingual Performance

In order to answer **RQ1** and **RQ2**, we compared the performances between a multilingual model and monolingual models trained specifically for each language.

4.1.1. Condition

The multilingual model was trained on all the data from the three languages mentioned above. Additionally, a monolingual model was trained separately for each language. Thus, the training data quantity of the multilingual model was three times that of each monolingual model.

The structure and parameters of the VAP model were the same for the multilingual and monolingual settings. The training parameters were as follows: The number of training epochs was 20, batch size was 8, learning rate was 3.63E-4, and weight decay was set to 0.001. We used the AdamW optimizer. We evaluated the test set using the model with the smallest loss on the validation set.

4.1.2. Test Loss Performance

As a basic evaluation metric, we assessed the average loss on the test set. The loss of interest for evaluation is L_{vap}, which was defined in Section 2. Table 1 shows the results. As can be seen, whereas the monolingual models work well when tested on the same language, they perform considerably worse when applied to another language. From this result, it is clear that the nature of voice activity projection differs across the three language datasets used in this study, and in order to make accurate predictions, it is necessary to train specific models for each language. However, the results of the multilingual model reveal that it can project voice activity with the same level of performance for all languages as the language-matched models.

Training data	Test data		
	ENG	MAN	JPN
English	**2.387**	3.401	2.956
Mandarin	2.839	**2.817**	3.098
Japanese	3.306	4.004	2.329
Multi (proposed)	2.396	2.832	**2.265**

Table 1: Test loss on cross-lingual performance

4.1.3. Turn Shift/hold Prediction

While the test loss reveals the general performance of the models, the numbers are hard to interpret. Thus, we also evaluated the applicability of a multilingual VAP model in the typical problem of predicting a *shift* vs. a *hold* in periods of mutual silence. This is an application that is similar to end-of-turn prediction in spoken dialogue systems. This evaluation is the same as the one used in previous research (Ekstedt and Skantze, 2022b). The task is to predict whether the preceding speaker and the following speaker were different (shift) or the same (hold) when a mutual silence longer than 0.25 seconds is observed. Note that the preceding and following utterances must be longer than one second. The value of p_{now} after 0.05 seconds into the start of the mutual silence is used to predict who the next speaker is.

The distribution of the shift/hold classes is summarized in Table 2. Since the nature of the dialogue varies depending on the language, the ratio of turn shifts to holds is also different. For example, in the Japanese data, there are fewer pauses (holds) and more concise utterances. The Mandarin data has the next largest imbalance, followed by English. In particular, in English, holds are approximately 10 times more frequent than shifts.

Although the evaluation metric used in previous research was the weighted F1 score, in this study, we used balanced accuracy to reduce the bias of class imbalance between languages. Furthermore, since the value of balanced accuracy would be 0.5 for random or majority-class prediction, it also has the advantage of being easily interpretable.

Table 3 shows the prediction results. The results are in line with those obtained using the test loss. Furthermore, when comparing the accuracy across languages, it is evident that Mandarin has the most predictable turn-taking patterns. From these results, we conclude that the monolingual models cannot be directly applied to other languages (**RQ1**), but that the multilingual model can be utilized as a generic turn-taking model for all three languages (**RQ2**).

Figure 5 illustrates output examples of the multilingual model in each language dataset. In these figures, both p_{now} and p_{future} are illustrated together with input waveforms colored with reference

Dataset	#Shift	#Hold	%Shift
English	1253	11432	9.9
Mandarin	718	1807	28.4
Japanese	1029	1371	42.9

Table 2: Distribution of samples for turn shift/hold prediction

Training data	Test data		
	ENG	MAN	JPN
English	**79.59**	68.64	59.43
Mandarin	65.31	84.49	59.72
Japanese	64.46	67.89	74.20
Multi (proposed)	77.16	**84.60**	**76.54**

Table 3: Cross-lingual turn shift/hold prediction performance (balanced accuracy [%])

VAD segments. In the English example (a), the turn shifts from the orange participant to the blue participant. During this turn shift, p_{now} exhibits a notably high predictive value before blue's speech. Additionally, during the pauses within each speaker's turn (holds), p_{now} and p_{future} correctly predict a continuation of each speaker's turn.

This same pattern is observed in the Mandarin example (b). In this case, we can see how p_{future} projects a turn-completion already before blue has stopped speaking. Furthermore, in the latter part of the Mandarin example, there is a short backchannel from blue, which is predicted by p_{now}. As this value is stronger than p_{future}, this illustrates how p_{now} and p_{future} can be used together to predict backchannels, but also for predicting that blue will not produce a longer utterance, after the onset of the backchannel.

The Japanese example (c) also illustrates that both turn shifts and holds can be predicted effectively. Note that there is a prolonged pause at the beginning of blue's turn-taking around the middle, during which p_{now} and p_{future} demonstrate an uncertainty as to who will be the next speaker. These types of situations are commonly observed in natural conversations, and are referred to as *self-selection* in the literature (Sacks et al., 1974). This kind of information could be utilized by spoken dialogue systems, allowing it to either take the turn or leave it to the user.

4.2. Language Identification

Next, to answer **RQ3**, we investigate the language identification ability of the multilingual model. Based on the observation that the multilingual model performed well in all three languages in the previous section, and that the monolingual models did not perform well for other languages, we hypothesized that the multilingual model is able to

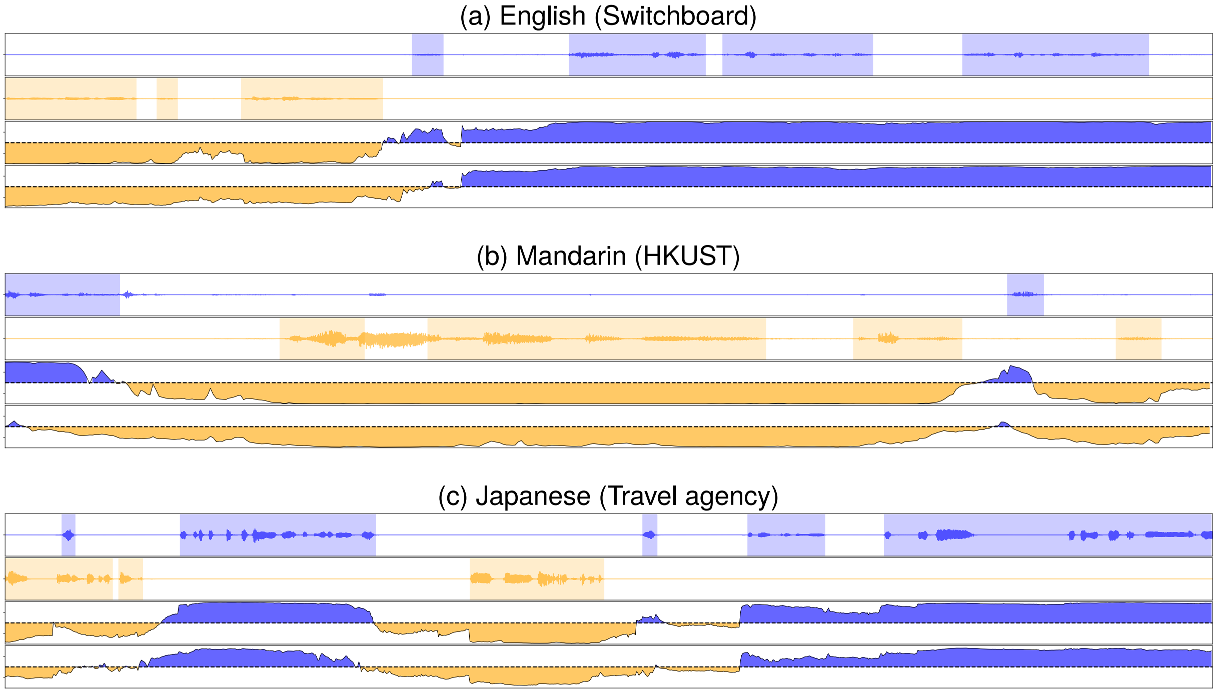


Figure 5: Output example of multilingual VAP in three languages (Top: English, Middle: Mandarin, Bottom: Japanese) - Each graph consists of, from top to bottom, input waveforms of both participants, near future voiced probability (p_{now}), and future voiced probability (p_{future}) among participants.

identify the language of the input speech and operates accordingly. To investigate this, we added another linear layer for language identification to the final layer of the VAP model, along with those for VAP and VAD. Since we are dealing with three languages, it becomes a three-class classification problem. Then, we added the cross-entropy-based language identification loss (L_{lid}) to the training loss as

$$L = L_{vap} + L_{vad} + L_{lid} \; . \tag{4}$$

We then trained this model from scratch in this experiment.

As a result, when the language identification accuracy was measured on the test set, it reached a weighted F1-score of 99.99%. In other words, the multilingual model is able to almost perfectly identify the language of the input speech.

We also wanted to see whether this added language identification loss would act as a multi-task loss, potentially improving the performance of the multi-lingual model. Table 4 reports the performance on the VAP test loss (L_{vap}) and the balanced accuracy of shift/hold prediction for turn-taking, for this new model compared to the previous model. As the results are very similar between models, we draw the conclusion that the model does not need help to learn to identify the language, but that it does so anyway implicitly.

Test data	Test loss (↓)		Shift/Hold (↑)	
	w/o LID	w/ LID	w/o LID	w/ LID
English	2.396	2.401	79.59	78.44
Mandarin	2.832	2.819	84.49	84.72
Japanese	2.265	2.341	74.20	75.82

Table 4: Performance with or without language identification (LID) multitask ("Shift/Hold" represents the balanced accuracy [%] on the turn shift/hold prediction task.)

4.3. Pitch Sensitivity

As mentioned in Section 1, prosodic information is an important factor in predicting turn-taking. To assess the model's reliance on prosodic information indirectly and to answer **RQ4**, we flattened the pitch of the input speech during the test phase (Figure 6) and measured the resulting performance degradation. A previous study conducted a similar test and found that pitch flattening had a minor overall impact on the model's performance, while being important in specific contexts of syntactic ambiguity (Ekstedt and Skantze, 2022a). In this study, we used Praat[2] to flatten the pitch, following the methodology of the aforementioned study. We then examined the performance changes before and after pitch flattening for the turn shift/hold prediction.

Table 5 presents the results for both monolingual

[2] https://www.fon.hum.uva.nl/praat/

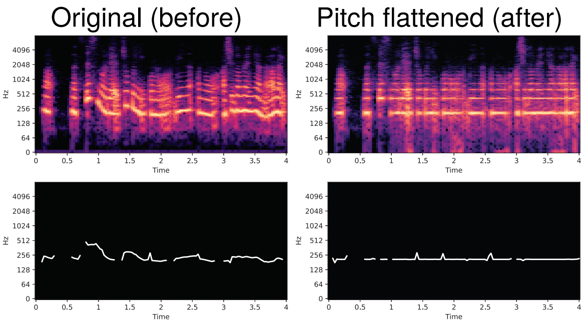


Figure 6: Input example of pitch flattening test (Top: Spectrogram, Bottom: Automatically estimated F0)

Test data	Mono	Multi
English	79.68 (+0.09)	76.28 (+0.12)
Mandarin	82.47 (−2.02)	82.30 (−2.30)
Japanese	72.83 (−1.37)	74.73 (−1.81)

Table 5: Turn shift/hold prediction performance (balanced accuracy [%]) with pitch flattening (difference against the case without pitch flattening)

and multilingual models. Both types of models exhibited similar trends, indicating that they rely on pitch information to a similar extent (**RQ4**). Furthermore, when analyzing the differences between languages, a decrease in accuracy of approximately two percentage units was observed in Japanese and Mandarin. In contrast, the change in accuracy for English was not significant. This discrepancy suggests that turn-taking in Japanese and Mandarin is more dependent on pitch cues, compared to English.

For Mandarin, together with the high accuracy in the shift/hold prediction experiment, our results suggest that pitch plays an important role for indicating turn-final position, which aligns with previous findings. While falling intonation is a universal turn-yielding cue in various languages (McCarthy, 1991), Mandarin, being a tonal language with lexical tones for individual words, may be able to provide richer pitch information for the model. For example, Levow (2005) investigated the interaction between tone and intonation and revealed that the pitch of turn-final words is nevertheless relatively lower than those in other positions in the utterance. They further show that intonation patterns can be used to train a classifier for determining turn-final syllables.

4.4. Effect of Audio Encoder

Finally, we also investigated the audio encoder used in the VAP model to answer **RQ5**. The encoder we used is CPC, which was also used in previous research on the VAP model. One potential limiting factor in a multilingual setting is that the model is trained on the English Librispeech dataset. Additionally, during the training of the VAP model, the parameters of this CPC are frozen. On the other hand, according to previous research, the English CPC model has been shown to be effective for phoneme recognition in other languages (Riviere et al., 2020). It should be noted that it is not straightforward to adopt any audio encoder for the VAP model, since the encoder needs to operate in a causal manner. This rules out models such as HuBERT (Hsu et al., 2021), which is bidirectional.

Test data	Test loss (↓)		Shift/Hold (↑)	
	CPC	MMS	CPC	MMS
English	2.396	2.421	79.59	77.67
Mandarin	2.832	2.841	84.49	82.09
Japanese	2.265	2.394	74.20	72.10

Table 6: Performance on comparison of audio encoders ("Shift/Hold" reports balanced accuracy [%] on the turn shift/hold prediction task.)

Still, we wanted to compare it with an audio encoder that is multilingual and not pre-trained on a specific language. Recently, Meta has released a multilingual wav2vec 2.0 model called Massively Multilingual Speech (MMS), which is pre-trained on data from 1406 languages (Pratap et al., 2023). Although the transformer layers of wav2vec 2.0 are bidirectional, the initial multi-stage CNN operates on a much smaller time window. Thus, we could adopt this initial part of the model for the VAP model, and compare the performance with CPC. Just like with CPC, this multi-stage CNN was frozen during the training of the VAP model.

The comparative results are reported in Table 6. Overall, MMS results show slightly lower performance compared to CPC. Thus, CPC seems to be more compatible with the task of the VAP model (**RQ5**). We also tried to train the entire model without freezing the audio encoder. However, both CPC and MMS showed a slight decrease in accuracy in doing so. Given the current size of the training dataset, there is a possibility that the model is overfitting.

5. Conclusion

In this paper, we have investigated the application of voice activity projection (VAP), a predictive turn-taking model for spoken dialogue, on multilingual data, encompassing English, Mandarin and Japanese. The results show that a monolingual VAP model does not work well when applied to other languages (**RQ1**). However, a multilingual VAP model (trained on all languages) shows comparable performance to monolingual models across all three language datasets (**RQ2**). Next, by incorporating an additional language identification task, we showed that the multilingual model can accurately identify the language of input audio (**RQ3**). We also investigated the model's sensitivity to pitch, by flattening the pitch of the input audio, Whereas the performance on English did not change by this perturbation, Japanese and Mandarin seem to be somewhat more dependent on pitch cues (**RQ4**). Finally, with respect to the audio encoder, we have found that the current pre-trained CPC model is better than the other alternative we have tried (**RQ5**).

Ethics Statement

This study does not involve any human subjects and we do not foresee any ethical consequences.

Acknowledgements

This work was supported by JST ACT-X JPMJAX2103, JST Moonshot R&D JPMJPS2011, JSPS KAKENHI JP23K16901, Riksbankens Jubileumsfond (RJ) P20-0484, and Swedish Research Council (VR) 2020-03812.

Bibliographical References

Patricia M Clancy, Sandra A Thompson, Ryoko Suzuki, and Hongyin Tao. 1996. The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of pragmatics*, 26(3):355–387.

Mark Dingemanse and Andreas Liesenfeld. 2022. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5614–5633.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283–292.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: A Transformer-based language model for predicting turn-taking in spoken dialog. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2981–2990.

Erik Ekstedt and Gabriel Skantze. 2022a. How much does prosody help turn-taking? Investigations using voice activity projection models. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 541–551.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice Activity Projection: Self-supervised learning of turn-taking events. In *INTERSPEECH*, pages 5190–5194.

Simon Garrod and Martin J Pickering. 2015. The use of content and timing to predict turn transitions. *Frontiers in psychology*, 6(751):1–12.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

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.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Real-time and continuous turn-taking prediction using voice activity projection. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*.

Yuichi Ishimoto, Takehiro Teraoka, and Mika Enomoto. 2017. End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous Japanese speech. In *INTERSPEECH*, pages 1681–1685.

Hua-Li Jian and Joyce Wu. 2011. Mandarin conversation: Turn-taking cues in exchange structure. In *International Congress of Phonetic Sciences (ICPhS)*, pages 970–973.

Kobin H Kendrick, Judith Holler, and Stephen C Levinson. 2023. Turn-taking in human face-to-face interaction is multimodal: Gaze direction and manual gestures aid the coordination of turn transitions. *Philosophical Transactions of the Royal Society B*, 378(1875):20210473.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and speech*, 41(3-4):295–321.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. In *INTERSPEECH*, pages 2658–2662.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2018. Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios. In *International Conference on Multimodal Interaction (ICMI)*, pages 78–86.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *International Conference on Multimodal Interaction (ICMI)*, pages 226–234.

Divesh Lala, Pierrick Milhorat, Koji Inoue, Masanari Ishida, Katsuya Takanashi, and Tatsuya Kawahara. 2017. Attentive listening system with backchanneling, response generation and flexible turn-taking. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 127–136.

Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):1–17.

Gina-Anne Levow. 2005. Turn-taking in Mandarin dialogue: Interactions of tone and intonation. In *SIGHAN Workshop on Chinese Language Processing (SIGHAN)*.

Wencke Liermann, Yo-Han Park, Yong-Seok Choi, and Kong Lee. 2023. Dialogue act-aided backchannel prediction using multi-task learning. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 15073–15079.

John K Local, John Kelly, and William HG Wells. 1986. Towards a phonology of conversation: turn-taking in tyneside english1. *Journal of Linguistics*, 22(2):411–437.

Ryo Masumura, Taichi Asami, Hirokazu Masataki, Ryo Ishii, and Ryuichiro Higashinaka. 2017. Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks. In *INTERSPEECH*, pages 1661–1665.

Michael McCarthy. 1991. *Discourse analysis for language teachers*. Cambridge university press.

Toshiki Muromachi and Yoshinobu Kano. 2023. Estimation of Listening Response Timing by Generative Model and Parameter Control of Response Substantialness Using Dynamic-Prompt-Tune. In *INTERSPEECH*, pages 2638–2642.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. 2023. Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation. In *International Conference on Human-Agent Interaction (HAI)*, pages 13–21.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint*. ArXiv:2305.13516.

Antoine Raux and Maxine Eskenazi. 2012. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing*, 9(1):1–23.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn taking for conversation. *Language*, 50(4):696–735.

Jin Sakuma, Shinya Fujie, and Tetsunori Kobayashi. 2023. Response timing estimation for spoken dialog systems based on syntactic completeness prediction. In *Spoken Language Technology Workshop (SLT)*, pages 369–374.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 220–230.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:101178.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences (PNAS)*, 106(26):10587–10592.

Nigel Ward and Wataru Tsukahara. 2000. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207.

Marcin Włodarczak and Mattias Heldner. 2016. Respiratory turn-taking cues. In *INTERSPEECH*, pages 1275–1279.

Language Resource References

Godfrey, John J and Holliman, Edward C and McDaniel, Jane. 1992. *SWITCHBOARD: Telephone speech corpus for research and development*. ISLRN 775-985-659-424-7.

Inaba, Michimasa and Chiba, Yuya and Higashinaka, Ryuichiro and Komatani, Kazunori and Miyao, Yusuke and Nagai, Takayuki. 2022. *Collection and Analysis of Travel Agency Task Dialogues with Age-Diverse Speakers*.

Liu, Yi and Fung, Pascale and Yang, Yongsheng and Cieri, Christopher and Huang, Shudong and Graff, David. 2006. *HKUST/MTS: A very large scale mandarin telephone speech corpus*. ISLRN 964-004-555-226-5.