The influence of prosody on turn-taking models at syntactically ambiguous places

Erik Ekstedt, Gabriel Skantze
Speech, Music & Hearing, KTH Royal Institute of Technology, Sweden
erikekst@kth.se, skantze@kth.se

Abstract

Turn-taking is a fundamental aspect of human communication and is the ability to organize turns, between the interlocutors, at appropriate locations throughout a conversation. In this work we investigate the influence of prosody on turn-taking using the recently proposed Voice Activity Projection model, which incrementally models the upcoming speech activity of the interlocutors in a self-supervised manner, without relying on explicit modelling of prosodic features, or specific annotations of turn-taking events. Inspired by psycholinguistic experiments we focus our analysis on single utterances containing syntactically ambiguous places, specifically designed to depend on prosody. We further investigate the implicit influence of prosody on the turn-taking model through prosodic manipulation of the speech signal.

Introduction

Turn-taking is the fundamental ability of humans to organize spoken interaction, i.e., to coordinate who the current speaker is, in order to avoid the need for interlocutors to listen and speak at the same time (Sacks, Schegloff, & Jefferson, 1974). A dialog can be viewed as a sequence of turns, constructed through the joint activity of turn-taking between the two speakers. A turn refers to segments of activity where a single speaker controls the direction of the dialog.

A common research question in phonetics, psycholinguistics and conversational analysis concerns the various cues (including speech, gaze and gestures) that humans use to detect or project turn-shifts (Duncan, 1972). When it comes to speech, a common distinction is made between the prosodic (non-lexical) and lexical (textual, syntactic, semantic) components of the speech signal. For example, (De Ruiter et al., 2006) argued, based on listening experiments, for the importance of syntactic information over intonation (pitch), while Bögels & Torreira (2015) showed that intonation is important when syntactic completion is ambiguous. However, such studies often require human listening experiments which are costly, anecdotal and constrained in time resolution and are therefore limited to small amounts of conversational contexts. An alternative approach is to use computational (Laskowski, Wlodarczak, & Heldner, 2019) to investigate what type of information they are sensitive to.

In conversational systems, turn-taking has traditionally been modeled using threshold policies which recognize silences longer than a chosen duration as transition-relevant places. Although these types of models are commonly used, it is well known that they are insufficient for modeling human-like turn-taking (Skantze, 2021). Studies of human-human conversation have shown that turns are frequently shifted with a gap of just 200ms (Levinson & Torreira, 2015), or even with a slight overlap. Thus, given that humans also need some time to prepare a response, it would be infeasible for humans to just use silence as a cue to turn-taking. Instead, it has been suggested that they are able to project turn completions already while the other person is speaking (Sacks et all., 1974; Garrod & Pickering, 2015; Levinson & Torreira, 2015). In addition, humans produce so-called backchannels (short feedback tokens such as "mhm") in a timely manner, often in overlap with the other speaker (Yngve, 1970).

Ekstedt & Skantze (2022) recently proposed Voice Activity Projection, **VAP**, which is a general, self-supervised turn-taking model. The model incrementally projects the future speech activity of the two speakers directly from raw audio waveforms. The model can be trained on lots of data, without human annotations, and is agnostic with respect to different types of speech information, as it does not depend on explicitly extracted features. This makes the VAP-model potentially suitable as a data-driven approach for investigating the role of prosody in turn-taking.

In this work, we train a VAP-model on a large dataset (Cieri, Christopher et al., 2004; Godfrey et al., 1992) of dyadic spoken interactions and evaluate it on specific turn-taking metrics, while augmenting the input audio to omit certain sources of prosodic information.

Background

Prosody refers to the non-verbal aspects of speech, including *intonation* (F0/pitch contour), *intensity* (energy), and *duration* (of phones and silences). It has been found to serve many important functions in conversation, including prominence, syntactic disambiguation, attitudinal reactions, uncertainty, topic shifts, and turn-taking (Ward, 2019). Studies on both English and Japanese have found that level intonation (in the middle of the speaker's fundamental frequency range) tend to serve as a turn-holding cue, whereas either rising or falling pitch can be found in turn-yielding contexts (Gravano & Hirschberg, 2011; Local et al., 1986; Koiso et al., 1998). When it comes to intensity, studies have found that speakers tend to lower their voices when approaching potential turn boundaries, whereas turn-internal pauses have a higher intensity (Gravano & Hirschberg, 2011; Koiso et al., 1998). Regarding duration and speaking rate, Duncan (1972) found a “drawl on the final syllable or on the stressed syllable of a terminal clause” to be a turn-yielding cue (in English). This is also in line with the findings of Local et al (1986).

When it comes to lexical information, a very strong cue to turn-taking is of course whether the utterance is syntactically or pragmatically complete (Ford & Thompson, 1996). Thus, even if prosodic cues can be found near the end of a turn-shift, it is not clear to what extent such cues provide additional information compared to lexical cues, or if they are redundant. In an experiment by De Ruiter et al (2006), subjects were asked to listen to a conversation and press a button when they anticipated a turn ending. The speech signal was manipulated to either flatten the intonational contour, or to remove lexical information by low-pass filtering. The results showed that the absence of intonational information did not reduce the subjects' prediction performance significantly, but that their performance deteriorated significantly in the absence of lexical information. From this, they concluded that lexical information is crucial for end-of-turn prediction, but that intonational information is neither necessary nor sufficient. Ekstedt and Skantze (2020) also found that it is possible to build fairly reliable turn-taking models using only lexical information.

However, it has also been argued that while lexical information is important for turn-taking, there are many cases where a phrase may be syntactically complete, but it is unclear whether the turn is in fact yielded or not (Ford & Thompson, 1996). To investigate this, Bögels and Torreira (2015) performed a similar experiment as De Ruiter et al (2006) but selected the stimuli so that they contained several syntactic completion points (e.g., "Are you a student / at this university?"), and where the intonation phrase boundary provided additional cues to whether the turn was yielded or not. They found that subjects indeed made better predictions with the help of intonation and duration.

Most previous attempts at modelling prosody in turn-taking have been limited in that they (I) only use instances of mutual silence for predicting turn shifts (and therefore do not model projection of turn completion), and (II) only use fairly superficial, hand-crafted features, such as the extracted pitch slope or pitch level right before the pause (Gravano & Hirschberg, 2011; Meena et al., 2014). Apart from the problem that such features might be too simplistic, they also typically require speaker normalization of pitch (Zhang, 2018).

In this work, we investigate various forms of turn-taking events (including projection of both turn shifts and backchannels). We also use a more agnostic modelling approach, using latent speech representations that are learned in a self-supervised manner and extracted from the raw waveform (Oord et al., 2018). If our model is indeed able to pick up relevant prosodic information from these representations, it means that we do not have to do any special prosodic feature engineering or speaker normalization.

Voice Activity Projection Model

Ekstedt and Skantze (2022) proposed a generic turn-taking model that does not predict specific turn-taking events at specific moments in time. Instead, the model is given the task of Voice Activity Projection (**VAP**), which means that it must incrementally predict the future voice-activity (**VA**) of each interlocutor in a dialog. The prediction target at each incremental step is defined by a window of 2 seconds containing the future VA for both speakers. The window is discretized into 8 separate bins (4 for each speaker) where each bin is assigned a value of one if more than half of its frames are active, to produce an 8-bit binary digit, corresponding to 256 unique classes.

The VAP model consists of an encoder which processes raw audio waveforms, along with the current VA-frame, VA^f, and a concise representation over its history, VA^h, to produce latent representations of a defined frame frequency, then fed into the predictor network. The predictor is a causal transformer (Vaswani et al., 2017) which processes the context available up until the current frame and outputs a probability distribution over the 256 VAP classes, see Figure 1. For further details we refer the reader to (Ekstedt and Skantze, 2022).

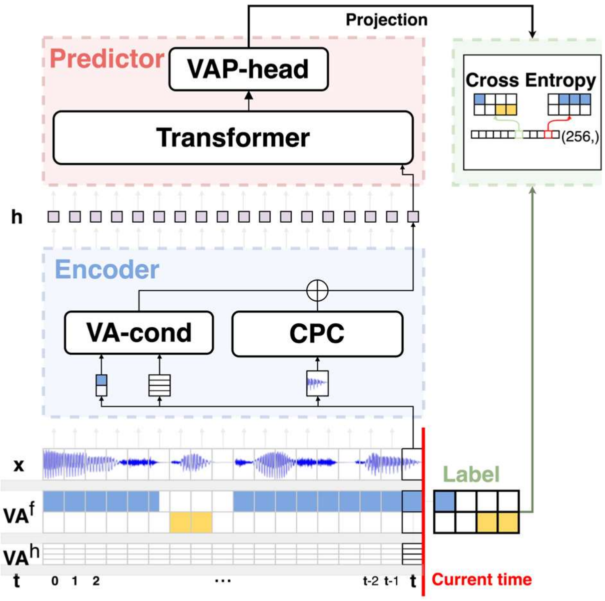


Figure 1. The VAP-model consists of an encoder and a predictor. The encoder processes raw waveforms and voice activity information to a latent representation used as input to the predictor. The predictor then outputs a probability distribution over all states defining the upcoming window of activity.

However, the model output can be difficult to utilize or interpret directly but Ekstedt and Skantze (2022) showed how it can be used to predict various turn-taking events as zero-shot classification tasks. The idea is to define subsets of classes that correspond to relevant transition states in the VA space then compare the probability mass over the given subset with another. The other subset can be the compliment (all other states), their opposite (the equivalent subset but from the other speaker's point of view) or any other subset. Inspired by Heldner & Edlund (2010) we select subsets over the distribution which corresponds to "clear" *gaps*, *pauses* and *overlaps-between*, which corresponds to VA transitions where the turn changes speaker.

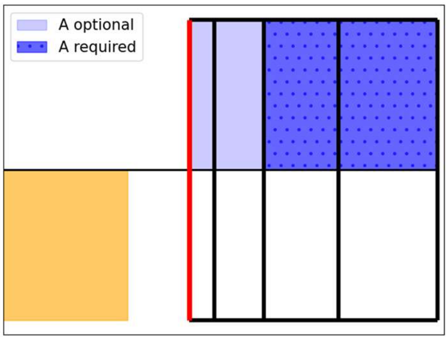


Figure 2. Template for the subset that encodes A (blue) being the next speaker at mutual silences. The red line indicates the current frame. Speaker B's bins must be inactive (white). Speaker A's last two bins must be active (dotted blue) and the first two are optionally active (light blue).

To determine the next speaker during segments of mutual silence we compare the subset corresponding to A as the only next active speaker, shown in Figure 2, with its opposite (i.e., only B is active). We constrain the subset by forcing the last bins to be active while the most immediate bins are optionally active. The subset changes slightly during segments of active speech where we allow the most immediate frames of the active speaker to optionally be active, shown in Figure 3. Here the probability mass is compared with the subset where only the current speaker is active, as described for mutual silences.

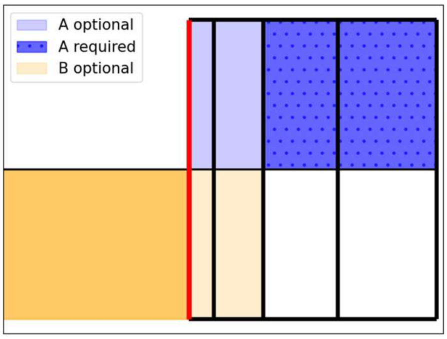


Figure 3. Template for the subset that encodes A (*blue*) being the next speaker during ongoing activity from speaker B (*yellow)*. The subset is the same as in Figure. 2 with the addition that the most immediate bins from speaker B may also be active (*light yellow)*. The red line indicates the current frame.

In other words, we force the model to choose one of the two speakers by comparing simplified subsets which clearly defines two possible outcomes. This enables the interpretation of the model output as probabilities associated with Shifts and Holds given knowledge about the last active speaker.

Training and Data

We train a VAP-model with a frame-level frequency of 50Hz (20ms frame size). We use a pretrained CPC-encoder (Rivière et al., 2020), kept frozen during training, to extract features from the combined, mono-channel, waveform of the two speakers. We use the combination of two dyadic conversational datasets, Switchboard (Godfrey et al., 1992) and Fisher-part-1 (Cieri et al., 2004), resulting in 8288 unique dialogs. We set aside a test set of 5% (of each dataset) and split the remaining dialogs into a 90/10 train/validation split used for training. We use the AdamW (Kingma & Ba, 2015; Ilya & Hutter, 2019) optimizer and an early stopping criterion on the validation loss with a patience of 10 epochs.

In order to investigate the role of prosody in the model's turn-taking predictions, we augment the input audio waveform of the test data in five ways to omit parts of the signal encoding for various prosodic features:

Low pass: the signal is low pass filtered by down-/up-sampling of the waveform like Weston et al. (2021). This effectively removes all high frequency phonetic information, while only the F0 and intensity contours are relatively intact. We use a cut-off frequency of 400Hz across all samples.

F0 flat: the intonation contour is flattened to the average F0 of each speaker and segment.

Intensity flat: The intensity contour is flattened to the average value of each speaker over all speech frames (as determined by the VA-features). We note that this augmentation is difficult to perform without including acoustic artifacts despite having access to speech boundaries given by the VA-features. Breaths become very loud and the gain inside smaller segments of silence is prominent.

Duration average: Each phone in a segment is scaled to the average duration, of that specific phone, across the dataset.

F0 shift: The intonation contour is shifted by 90% of the original value for each speaker over each active speech segment. This should (in theory) not affect the turn-taking predictions. However, we include this augmentation to verify that the augmentations themselves does not have a too strong effect (e.g., through artifacts).

All code is implemented in Python using the PyTorch (Paszke, et al., 2019), PyTorch-Lightning (Falcon, 2020) and Wandb (Biewald, 2020) libraries for machine learning and Praat (Yannick, Thompson, & De Boer, 2018; Boersma & Weenink, n.d.) for augmentations.

Aggregate Turn-shift Evaluation

We evaluate the model on the test-set given to the turn-shift interpretations described above. The evaluation events are the same as described in Ekstedt and Skantze (2022), namely **Hold/Shift** and **Pred-Shift**, which automatically extracts frames of turn-shifts at mutual silences and during active speech. We evaluate on the original audio as well as augmented versions with the exception for *Duration average*, given that we do not have access to phone aligned annotations of the datasets. The aggregate performance is visualized in Figure 4.

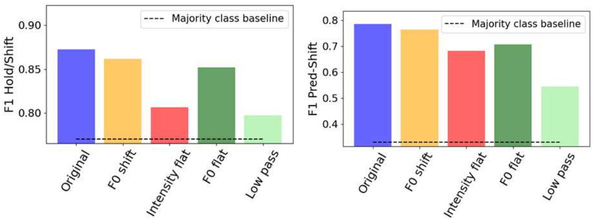


Figure 4. Aggregate results over the dialog test-set using the original and augmented waveforms. **Left**: Hold vs Shift at mutual silences. **Right**: Shift prediction at ongoing speech. Note the difference in scale of the y-axis.

We note that the Shift/Hold metric is highly imbalanced, containing a substantially larger number of holds, indicated by the high (.77) majority class baseline. The Shift-prediction metric is balanced by design, resulting in a lower baseline value (.33). The least intrusive augmentation is, as expected, the *F0 shift* transformation. However, the artifacts introduced still seem to have some effect on the models. The *Low pass* augmentation have the most significant impact on performance for over both tasks. This augmentation basically omits all information other than the F0 and intensity contours and indicates that the model do rely on more complex cues to predict the next speaker. The second most impactful augmentation is *Intensity flat*, which indicates, in accordance with the turn-taking literature in general, that shifts are preceded by changes (arguably drops) in the intensity contour of the current speaker. We note that it seems slightly more important for the pred-shift task. Interestingly, *F0 flat* had the least negative effect, which is surprising, given that pitch seems to be the most frequently used prosodic cue in computational turn-taking models.

Utterance-level Analysis

While the analysis above gives an overall metric on how important prosody is, it has been hypothesized that prosody is especially important when the semantic/pragmatic completion is ambiguous, as discussed in Background. To focus their analysis on such situations, Bögels and Torreira (2015) constructed question templates where a short and a long version, sharing initial lexical information, were recorded through scripted interviews (in Dutch). As an example, a short/long question pair "did you drive here?" and "did you drive here this morning?" contain the same initial words up to a common completion point (after the word "here"), which we will refer to as the *short completion point*, **SCP**. Note that for the listener (or the model) to predict a turn-shift towards the end of the short utterance, but not at the corresponding place in the long utterance, it must rely on prosody. Through listening experiments, where the participants are asked to press a button when they expect a turn shift, Bögels and Torreira (2015) found that the reaction time was indeed much faster after the short version, than after a long version cut after the SCP.

For our experiments, we created a similar set of 9 long/short utterance pairs in English using the Google TTS service and produced 10 versions of each long/short pair using 5 male and 5 female voices. The phrases are listed in Table 1. An example of such a pair, along with the model's *Shift-prediction*, is visualized in Figure 5. As can be seen in the figure, the model correctly assigns a high probability to Hold until towards the end of each utterance, where it changes to Shift. This clearly illustrates the model's ability to project turn shifts before the utterance is complete, and before the large rise in final pitch has happened. In addition, we see how the model makes a clear distinction between the two utterances at the short completion point (SCP), where it predicts a Hold for the longer variant. This illustrates that the model is indeed sensitive to prosody, as that is the only information that is different up until that point.

Table 1. The 9 phrases used in the utterance level analysis.

Short	Long
Are you a student	… here at this university
Do you study psychology	… here at this university
Are you a first year student	… here at this university
So do you play basketball	… on Thursdays
Have you participated in any experiments before	… here at this university
Do you live by yourself	… or with someone else
So you work on the side	… in a supermarket in addition to your studies
Did you come here by bike	… this morning
Did you drive here	… this morning

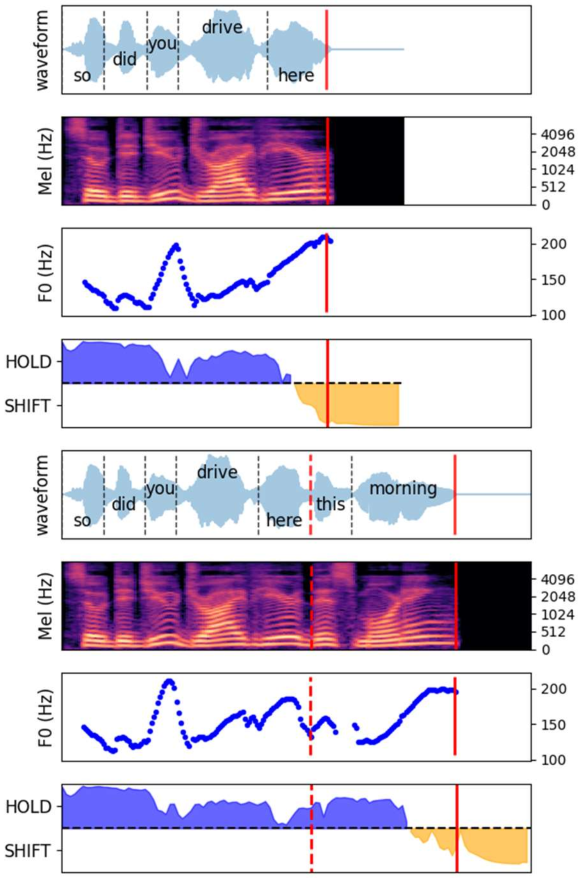


Figure 5. A short/long phrase pair. The plots show the waveforms, mel-spectrograms, F0 contours and the model assigned Shift/Hold comparison, for the short and long version respectively. The blue color in the bottom plots indicate a probability over 50%. The SCP is shown as a red dashed line for the long utterance. The red lines show the end time of the last word in each utterance.

Since we rely on artificially generated utterance pairs, we are uncertain to what extent they reflect similar prosodic patterns as those generated by humans. Therefore, we perform a similar analysis of the phrases as Bögels and Torreira (2015), by measuring the duration and maximum F0 frequency over the last syllable of the short completion point. In their analysis, they showed that longer duration and a higher rise in F0 are associated with the end of a turn, separating the measures at the SCP of the short phrase from the long. We obtain similar distributions from four out of the nine phrases but note that the remaining ones are not as easily separable, showing more uniform distributions over the duration dimension, as shown in Figure 6. However, from listening to the phrases, we still consider all recordings natural enough to be included in our analysis.

We compare the performance of the VAP model on the short and long versions of each phrase to investigate whether it can recognize the prosodic differences and correctly predict the short completion point as either a Hold (long phrase) or a Shift (short phrase). In addition to the original recordings, we include evaluations of the

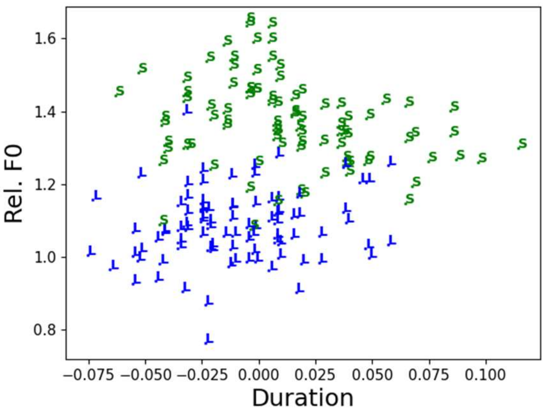


Figure 6. Duration and maximum relative F0 over the last syllable at the "short completion point" for (L)ong and (S)hort versions of the synthesized voices. The x- and y-axis correspond to mean-shifted duration and relative F0 peak.

performance on the augmented versions to investigate whether any specific augmentation changes the predictions of the model more than the other. The model output for the long version of the phrase "Are you a student here at this university?", given various augmentations, is visualized in Figure 7-8.

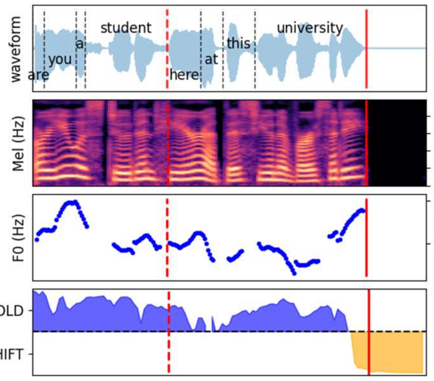
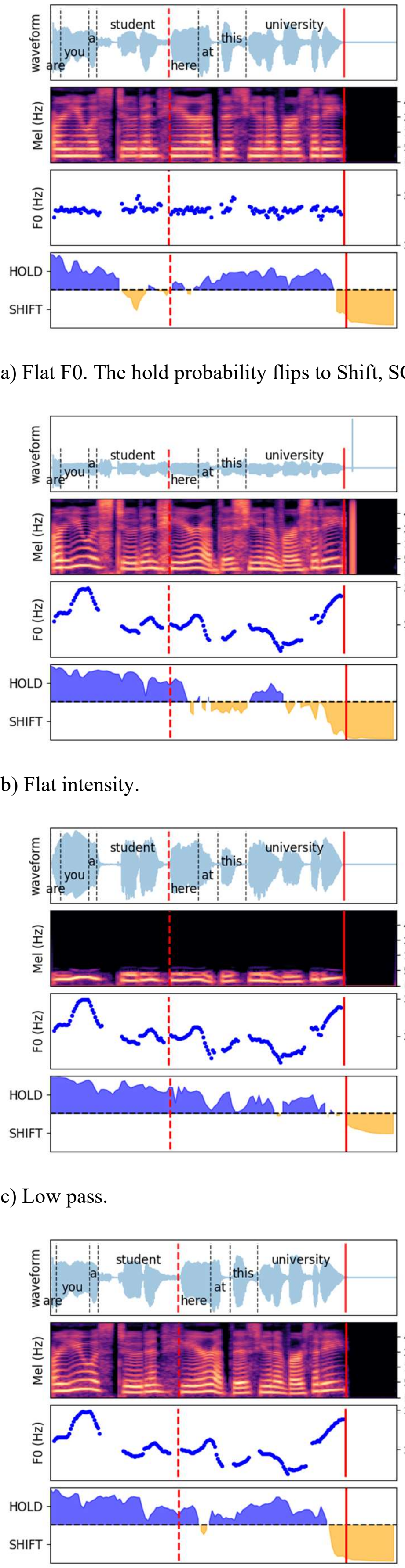


Figure 7. Original waveform of the long phrase "are you a student here at this university?".

Inspection of the original performance in Figure 7 indicates that the model is sensitive to prosodic information and assigns a higher Hold probability at the SCP located on the word "student". However, for the *F0 flat* augmentation, in Figure 8a, we note that the model flips and assigns a higher Shift-probability at the SCP, which indicate that if the dynamics of the F0 contour is omitted, the model cannot recognize that the speaker will continue to speak. Interestingly, the *Intensity flat* augmentation also effects the output of the model, but after the SCP, shown in Figure 8b. Here, the model does have access to the F0 contour and correctly assigns a larger Hold-probability at the SCP, but then changes prediction to indicate that a Shift is probable following the word "here". As a final note, the *Low pass* augmentation, which filters out all phonetic information while keeping both the intensity and F0 contour, does produce predictions close to that of the original audio, while being slightly less certain of a Shift after the entire utterance is completed, as seen in Figure 8c. The average duration augmentation, in Figure 8d, seems to have a minute effect on this phrase where



(a) Flat F0. The hold probability flips to Shift, SCP

(b) Flat intensity.

(c) Low pass.

(d) Average duration.

Figure 8. Augmented versions of the long phrase "are you a student here at this university?".

the only qualitative difference from the original occurs after the SCP at the word “here”. The slightly longer phones do trigger a short *Shift* segment hinting towards a model sensibility for a “drawl” which is an elongation of phones at the end of turns.

To get an aggregate evaluation of the model across all phrases, we define three regions in each utterance, up until the SCP point (for both long and short phrases), namely **hold**, **predictive** and **reactive**, and measure the average Shift probability predicted by the model in those regions. The *hold* region covers the beginning of the utterance until the *predictive* region starts 200ms before the SCP. The *predictive* region continues until the very last frame, referred to as the *reactive* region, of the SCP where the entire word has been spoken. The model should produce a low shift probability on the *hold* region on both the long and short versions of the phrase. Over the short phrase the shift probability should increase across the *predictive* and *reactive* regions indicating that the model correctly predicts the SCP as the end of the utterance. However, the opposite is true over the long versions where the model should predict a hold. The aggregate model performance over all phrases is visualized in Figure 9.

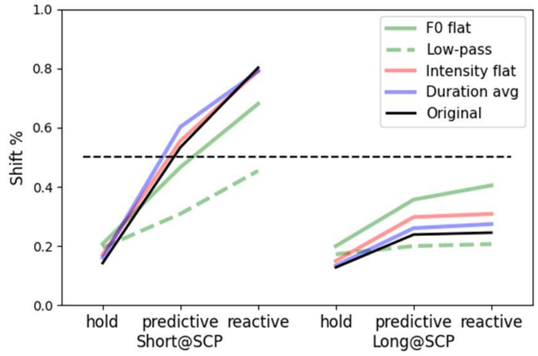


Figure 9. Shift probabilities over the three regions on the short completion point over all phrases.

The left part of Figure 9 displays the average Shift probabilities for the points on the SCP for the short phrases (Short@SCP) which preferably should start low and rise consistently. The right part of the figure shows the corresponding performance but on the long phrases (Long@SCP) and should be consistently low, indicating that the speaker will continue their turn. Looking at the non-augmented signal (Original), and comparing the left and right figures, we see that the model is indeed sensitive to prosody, confirming the anecdotal observation from Figure 5. The *Low pass* augmentation clearly hinders the model from predicting a Shift, indicating that pitch and intensity in themselves are not enough. Among the other augmentations, *F0 flat* seems to have the largest negative effect, which confirms that intonation is important for disambiguating turn completion when lexical information is not enough. Duration seems to be less important, which aligns with the observation in Figure 6.

Conclusion and Discussion

In this work we train a general computational model of turn-taking, provide analytical methods suitable for evaluating their performance on turn-shift classification, and investigate how they utilize prosodic information of the speech signal. We investigate the model’s reliance on prosody by extending psycholinguistic experiments designed to measure the effect of prosody for turn-taking in human subjects.

We apply specific prosodic augmentations to the input signal and show a deterioration of performance over two forms of turn-shift tasks, namely predictions at ongoing speech and during mutual silences. The performance on the dataset of human-human dyadic conversation is less effected by prosodic manipulation than the specifically designed phrases. We note that omitting phonetic information, through the *Low pass* augmentation of the signal, has the largest impact on both the aggregate- and utterance-level evaluation. This is not surprising given that it is the most intrusive transformation of the signal but does indicate that VAP models do use phonetic information to model future activity. However, it is interesting that on the aggregate-level analysis a flattening of the F0 contour does not seem to drastically effect performance, instead intensity seem to play a more important role.

Even more convincing are perhaps the specific comparisons of the model’s ability to predict Shift vs Hold at syntactic completion points, where the lexical information is identical. This task requires access to the prosodic dynamics of the signal and should be impossible to distinguish based on lexical information alone. It is interesting to note that on the syntactically ambiguous completion points F0 seem to play a more important role than that of intensity, showing the opposite effect of the aggregate-level analysis.

Overall, we show that all models are most sensitive to the *low-pass* augmentations, indicating that phonetic information is important for turn-taking in general. We note that intensity is at least as important as pitch when applied to actual human long-form conversations, but that pitch plays a more important role for the disambiguation at syntactically equivalent completion points. Interestingly, we note that the importance of duration plays a less important role, indicating that the F0-contour is the most reliable cue in the presence of lexical ambiguity.

We define an automatic evaluation task using TTS generated utterances, inspired by a psycholinguistic experiment designed for humans, which enables the possibility to perform coherent evaluation over large amount of data not reliant on expensive human evaluations. This form of evaluation of computational models could aid in understanding, interpreting and design of deep-learning models while at the same time serve to increase cooperation between deep-learning researchers and the linguistic, psycholinguistic and conversational analysis communities.

Acknowledgements

This work was supported by Riksbankens Jubileumsfond (RJ), through the project *Understanding predictive models of turn-taking in spoken interaction* (P20-0484), as well as the Swedish Research Council, through the project *Prediction and Coordination for Conversational AI* (2020-03812).

References

Biewald, L. (2020). ExperimentTracking with Weights and Biases. Retrieved from https://www.wandb.com/

Boersma, P., & Weenink, D. (n.d.). Praat: Doing phonetics by computer. Retrieved 05 13, 2022, from http://www.praat.org/

De Ruiter, J., Mitterer, H., & Enfield, N. (2006, 09). Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation. Language, 82, 515-535. doi:10.1353/lan.2006.0130

Duncan, S. (1972, 08). Some Signals and Rules for Taking Speaking Turns in Conversation. Journal of Personality and Social Psychology, 23, 283-292. doi:10.1037/h0033031

Ekstedt, E., & Skantze, G. (2022). Voice Activity Projection: Self-supervised Learning of Turn-taking Events. arXiv. doi:10.48550/ARXIV.2205.09812

Falcon, W. (2020). PyTorchLightning/pytorch-lightning. doi:10.5281/zenodo.3828935

Ford, C., & Thompson, S. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. Schegloff, & A. Thompson, Interaction and grammar (pp. 134-184). Cambridge: Cambridge University Press.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. Computer Speech & Language, 601-634.

Ilya, L., & Hutter, F. (2019). Decoupled Weight Decay Regularization. International Conference on Learning Representations, ICLR.

Kingma, D., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations, ICLR.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. Language and Speech, 295-321.

.org/Y18-1095

Laskowski, K., Wlodarczak, M., & Heldner, M. (2019). A Scalable Method for Quantifying the Role of Pitch in Conversational Turn-Taking. Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (pp. 284-292). Stockholm: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-5934

Local, J., Kelly, J., & Wells, W. (1986). Towards a Phonology of Conversation: Turn-Taking in Tyneside English. Journal of Linguistics, 411-437.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lere, A. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024-8035). Curran Associates, Inc. Retrieved from {https://pytorch.org/

Sacks, H., Schegloff, E. A., & Jefferson, G. D. (1974). A simplest systematics for the organization of turn-taking for conversation. Language, 50, 696-735.

Skantze, G. (2021). Turn-taking in Conversational Systems and Human-Robot Interaction : A Review. Computer Speech & Language (p. 101178). Elsevier Ltd. doi:10.1016/j.csl.2020.101178

Ward, N. (2019). Prosodic Patterns in English Conversation. Cambridge University Press. doi:10.1017/9781316848265

Yannick, J., Thompson, B., & De Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics, 71, 1-15. doi:https://doi.org/10.1016/j.wocn.2018.07.001

Yngve, V. H. (1970). On getting a word in edgewise. Chicago Linguistics Society, 6th Meeting, 1970, (pp. 567-578). Retrieved from https://ci.nii.ac.jp/naid/10026753338/en

Zhang, J. (2018). A Comparison of Tone Normalization Methods for Language Variation Research. Association for Computational Linguistics. Hong Kong. Retrieved from https://aclanthology