Toward Real Time Word Based Prosody Recognition

Alex Tilson and **Frank Förster**
Robotics Research Group
School of Physics, Engineering and Computer Science
University of Hertfordshire
AL10 9AB, United Kingdom
{a.tilson, f.foerster}@herts.ac.uk

Abstract

Prosodic salience is a heuristic based on word-level prosody in child-directed speech that is thought to serve as a cue for attentional focus. It has been used in the context of robotic language acquisition to extract the contextually most relevant words from a human tutor's speech to ground them in a robot's sensorimotor data. However, the pipeline for performing word-based prosody-recognition operated in a semi-automatic manner and required substantial manual effort. We describe our efforts to automate the existing pipeline by including real time prosody recognition, and a modern speech recognition and forced alignment model. The intention is to enable its use in real time for human-in-the-loop robotic language acquisition and other socially driven forms of online learning.

1 Introduction

Prosodic salience is a measure calculated from a speech signal's pitch, energy, and duration features, and can be used to identify the most relevant words of an utterance produced by caregivers in child-directed speech.

This heuristic has demonstrated use in robotic language acquisition (Saunders et al., 2011, 2012), and can facilitate a more effective language learning process for robots, drawing on insights from how human children acquire language.

It has been used as part of the ITALK project (Broz et al., 2014) to learn the names of, and interactions with, objects based on human tutors' linguistically unconstrained speech when trying to teach the robot the names of various objects after having been told to speak to the robot as if it were a 2-year old child.

Further research was performed by (Förster et al., 2011; Förster et al., 2019) demonstrating that negation words, such as "no", are prosodically salient which may explain why it is typically amongst the first 10 words in English-speaking children's early active vocabularies (Fenson et al., 1994).

While the previous work shows prosodic salience to be useful for word-level language acquisition in developmental robotics, it may have a wider potential in speech interfaces. For instance, it might be used within dialogue systems in generating different responses depending on whether some word was produced meekly or with strong intonation - think of the difference between a meekly uttered "no" and a vehemently shouted one.

However, for word-based prosody recognition, features must be aligned accurately to the correct segment of the speech signal that is representative of the word. For prosodic salience, this meant that a large quantity of manual effort was required in the past from human transcribers marking word boundaries (e.g. Saunders et al., 2011, 2012; Förster et al., 2011; Förster et al., 2019). Hence for any meaningfully large corpus to be processed by this method, it would need to be scaled up by automating the alignment process with speech processing methods and speech recognition models.

The present paper describes our efforts to automate the existing semi-automatic prosody-processing pipeline.

2 Background

2.1 Child-Directed Speech and Language Acquisition

Child-directed and infant-directed speech (CDS and IDS respectively) are marked out by a number of modifications compared to adult-directed speech, that have been hypothesized to be conducive to human language acquisition. While not all of these modifications are present in all languages, they typically include an overall higher pitch, exaggerated intonation contours, a focus on topics relating to physically co-present objects or events, and important words being placed at an utterance-final

Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning, pages 62–67
October 14–15, 2024. ©2024 Association for Computational Linguistics

position (Clark, 2009, chap. 2). Moreover, objects words have been observed to be pronounced relatively loudly (Saxton, 2017, chap. 4). While most of these observations characterise CDS and IDS on a general level, Soderstrom (2007) hypothesises that in IDS some of these acoustic modifications are performed on a word-level to aid the infant in both segmenting an utterance and singling out a target word.

In the context of robotic language acquisition and for the purpose of symbol grounding, based on the aforementioned features of CDS, Saunders, Lehmann, Sato, and Nehaniv (2011), operationalised word-based prosodic salience (cf. section 3). This was done to identify and extract prosodically salient words from an utterance produced by a human tutor when speaking to a child-like humanoid robot. Here, the prosodic salience of a word is identified as the product of a word's normalised pitch, energy, and duration values.

2.2 Human-in-the-Loop Real-time Reinforcement Learning

Senft et al. (2019) created an implementation for a reinforcement learning agent that learns from social feedback in an education setting. The reward signal the robot learned from was in the form of corrective feedback via a human manually pressing buttons to reward, punish, or manually initiate, actions. Because the teacher must consciously provide explicit feedback to the robot, their workload did not sufficiently decrease over time.

Belpaeme et al. (2018) express that the use of explicit signals in these cases acts as a proxy for naturally expressed implicit social signals. As some of these signals are typically embedded within speech, they contend that speech processing technology presented a bottleneck in their study preventing them from using such implicit speech-based social signals.

Prosodic salience is an example of such implicit social signals and similarly suffers from this bottleneck because of its reliance on the temporal alignment between the lexical level and the audio signal.

2.3 Forced Aligners

Forced alignment (FA) is the process of aligning a transcript to an audio signal. The traditional approach to forced alignment makes use of Hidden Markov Model (HMM) based automatic speech recognition pipelines, where statistical methods are used to model the probability distributions of phonetic units and to align the audio with the text.

Whilst traditional speech recognition models have largely been surpassed by attention based models such as (Baevski et al., 2020) and (Radford et al., 2023), attention based forced aligners haven't improved performance as significantly. For instance, NeuFA (Li et al., 2022) only marginally improves on the HMM based Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) and WhisperX (Bain et al., 2023) simply performs worse.

3 Methods

3.1 Prosodic Salience Pipeline

The prosodic salience estimation pipeline (Saunders et al., 2011) is as follows:

1. Transcribe the speech signal
2. Align the transcription
3. Split words into groups of utterances
4. Estimate the mean pitch and mean energy features of each spoken word
5. Estimate salience
6. Create a lexicon using the most salient words of each utterance
7. Ground sensorimotor experience with lexical units.

Originally steps 1, 2, and 3 were performed semi-automatically with human correction, with transcription alignment comprising the majority of manual effort from human transcribers. To automate these processes, we used Deepgram's Nova model (Deepgram, 2024) through an API call, which automatically produces a transcript, word boundaries, and utterance boundaries.

Additionally, step 4 relied on the Prosodic Feature Extraction Tool (PFET) (Huang et al., 2006), to calculate the relevant pitch, duration and energy features. However, as it was built on top of Praat (Boersma and Weenink, 2024), it was designed as an analysis tool, and has limited capabilities for full automation and real time execution.

To automate this process, we use OpenSMILE (Eyben et al., 2010) which is a highly configurable open source toolkit for signal processing and audio feature extraction. Comparitively, it has real time execution capabilities, and can be fully automated.

The calculation for estimating prosodic salience (or step 5), which is independent of the pipeline, is as follows: for a given word W_i of an utterance U, with $U = [W_1, \ldots W_n]$, the word-based mean pitch, energy, and duration are scaled with respect to the maximum word-based mean value of the respective measure within U (p = pitch, e = energy, d = duration, s = prosodic saliency).

$$\hat{p}(U) = max(\{p(W_i) \mid W_i \in U\}) \quad (1)$$

$$\hat{e}(U) = max(\{e(W_i) \mid W_i \in U\}) \quad (2)$$

$$\hat{d}(U) = max(\{d(W_i) \mid W_i \in U\}) \quad (3)$$

$$s(W_i) = \frac{p(W_i)}{\hat{p}(U)} \times \frac{e(W_i)}{\hat{e}(U)} \times \frac{d(W_i)}{\hat{d}(U)} \quad (4)$$

For single word utterances, if the word's pitch, energy, or duration are larger than the first standard deviation of pitch, energy, and duration for the whole interaction session, then the word is also marked as salient.

3.2 Analysis

Step 1: Test of partially-modified pipeline using OpenSMILE To test the suitability of OpenSMILE to act as replacement for PFET, the modified pipeline using OpenSMILE was compared against the original pipeline using PFET by running them on robot-directed speech (RDS) corpus of Förster et al. (2019). For this corpus both manual transcriptions and word boundary time stamps are available, such that no additional method to detect word boundaries such as Deepgram is needed. Executing both pipelines yielded two sets of prosodically salient words whose frequencies we subsequently compared using Kendall rank correlation test.

Step 2: Test of fully-automated pipeline using both OpenSMILE and Deepgram The fully-automated pipeline was tested using Deepgram's Nova model which can generate both speech transcripts and word boundary time stamps. Using this pipeline, we generated a speech aligned transcript for the Newman-Ratner CDS corpus (Newman et al., 2016). This corpus was chosen due to the similarity of the two scenarios within which both the Förster and Newman-Ratner corpora were recorded. For reasons of data protection we were not allowed to upload the Förster corpus into the cloud-based Deepgram, hence the need for the Newman-Ratner corpus. The utterance boundaries and word alignments generated by Deepgram were then used with the prosodic features generated by OpenSMILE to calculate the duration, mean pitch, and mean energy values for each word, followed by calculating their prosodic salience. Subsequently the prosodically most salient words, one per utterance, were extracted for this corpus, and table of word frequencies created (cf. section 4). Upon reviewing the extracted words, we noticed an unexpected absence of object labels, which are known to occur frequently in child-directed speech and we would expect to be prosodically salient, as seen in the Förster corpus. This necessitated additional analyses to investigate the cause of the dissimilarity between the two corpora.

Follow-up Analysis: Forced Aligners on RDS Listening to the selected section of the audio recordings of the Newman-Ratner corpus made it clear that the fully-automated pipeline had failed to pick out the prosodically most salient words. After verifying the correctness of the speech transcripts generated by Deepgram, two potential error sources were identified: (1) a failure of OpenSMILE to correctly calculate the different prosodic feature values, for example due to noise or poor audio quality, and (2) a failure of Deepgram to correctly determine the word boundaries, leading to a misalignment of transcript and audio recording.

Hence Deepgram's alignment accuracy was tested using a test audio file from the Förster corpus and by comparing Deepgram's word boundaries to the human-generated baseline. The file was 193 seconds long, consisting of 181 words, of which only 141 were used, as they were a part of an utterance which contained a saliently predicted word and therefore the most likely to affect the results. To account for cases where more than one word was produced by Deepgram, word alignments were paired based on the closest match of start and end timestamps. Algorithm 1 was used to quantify the degree of misalignment.

Algorithm 1 *Overlap Function* a and b are time intervals under comparison, specifying word boundaries as tuples, with a: ground-truth, and b: other boundaries (here: generated by Deepgram).

```
function OVERLAP(a, b)
    $a\_len \leftarrow |a[1] - a[0]|$
    $b\_len \leftarrow |b[1] - b[0]|$
    $overlap \leftarrow \min(a[1], b[1]) - \max(a[0], b[0])$
    $missing \leftarrow a\_len - overlap$
    $extra \leftarrow b\_len - overlap$
    return $(overlap, missing, extra)$
end function
```

The overlap function calculates the overlap, missing length, and extra length between two intervals a and b.

4 Results

4.1 Prosodic Feature Extraction

Step 1 The outcome of the analysis performed in Step 1 is depicted in Fig. 1. Shown are the relative frequencies of the top 10 most frequent prosodically salient words of the Förster corpus for both pipelines. The Kendall rank correlation test yielded a $\tau_B = 0.86$ ($p <= .001$), indicating a large correlation.

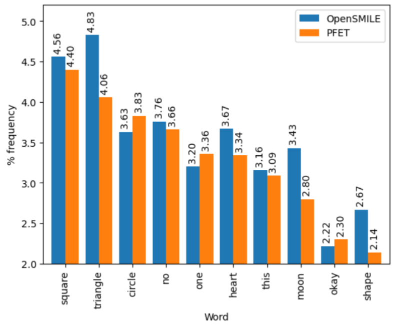

Figure 1: Frequency of the 10 most frequent prosodically salient words by pipeline. The pipeline is identical besides the prosodic feature extraction method. Methods compared are the originally used Chen and Harper's prosodic feature extraction tool (PFET) vs OpenSMILE.

Step 2 Table 1 depicts the 10 prosodically most salient words as detected by the fully-automated pipeline. Objects labels, dominating the list of most-frequent prosodically salient words in Förster's RDS corpus are suspiciously missing here, indicating a problem with the prosody detection.

Word	Freq.	Word	Freq.
oh	1367	baby	457
yeah	1032	look	436
you	700	that	339
okay	669	what	240
no	665	see	238

Table 1: 10 most frequent prosodically most salient words of the Newman-Ratner corpus by frequency of occurrence over all participants as output by the fully-automated pipeline.

4.2 Forced Aligners on Robot Directed Speech

Table 2 shows the total overlap, missing, and extra sections between the baseline word alignment and the one generated by Deepgram when run on the test audio file from the RDS corpus. Numerically, the missing and extra parts of audio account for nearly the same portion as overlap. This is catastrophic for prosodic salience estimation, as word level prosody data can change frequently, and nearly half of it is either erroneous or missing.

Category	Time (seconds)
Overlap	34.021
Missing	14.366
Extra	15.678

Table 2: Totals for Overlap, Missing, and Extra, segments of audio for Deepgram's prediction compared to human aligned. Overlap represents the total time in seconds where the time ranges agree. Missing represents the portions of audio where the prediction undershoots the word. Extra represents the portions of audio where the predicted region overshoots the word.

5 Discussion

Our results indicate that word boundary detection as performed by forced aligners, still remains an open problem when applied to child-directed speech and with respect to word-based prosody detection. We observed that once a boundary detection error occurs within an utterance, this type of error frequently propagates to the boundaries of subsequent words in that utterance. This subsequently renders word-based prosody detection difficult to impossible. However, given a correct set of correct word boundaries, current automatic prosody feature extraction tools such as OpenSMILE appear to perform sufficiently well when compared semi-automatic prosody processing methods involving tools such as PRAAT. Because traditional FA performs poorly in non-standard domains, settling for a hybrid usage of HMM and attention models for speech alignment appears to be insufficient. Purely attention based forced alignment models hold some promise for improvement.

Future Work Elsner and Ito (2017) posit that forced aligners perform poorly on CDS due to its atypical phonetics, resulting in what they call "catastrophically aligned words". In their work, a Kaldi forced aligner was adapted to CDS by treating it as a domain adaptation problem. We hence intend to tune NeuFA (Li et al., 2022) and similar attention-based aligners to chosen CDS and RDS corpora to adequately adapt it to the respective domains.

References

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

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

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

Paul Boersma and David Weenink. 2024. Praat: doing phonetics by computer [computer program]. Retrieved 31st May 2024 from http://www.praat.org/.

Frank Broz, Chrystopher L. Nehaniv, Tony Belpaeme, Ambra Bisio, Kerstin Dautenhahn, Luciano Fadiga, Tomassino Ferrauto, Kerstin Fischer, Frank Förster, Onofrio Gigliotta, Sascha Griffiths, Hagen Lehmann, Katrin S. Lohan, Caroline Lyon, Davide Marocco, Gianluca Massera, Giorgio Metta, Vishwanathan Mohan, Anthony Morse, Stefano Nolfi, Francesco Nori, Martin Peniak, Karola Pitsch, Katharina J. Rohlfing, Gerhard Sagerer, Yo Sato, Joe Saunders, Lars Schillingmann, Alessandra Sciutti, Vadim Tikhanoff, Britta Wrede, Arne Zeschel, and Angelo Cangelosi. 2014. The ITALK Project: A Developmental Robotics Approach to the Study of Individual, Social, and Linguistic Learning. *Topics in Cognitive Science*, 6(3):534–544.

Eve V. Clark. 2009. *First Language Acquisition*. Cambridge University Press, Cambridge, UK.

Deepgram. 2024. Deepgram Voice AI: Text to Speech + Speech to Text APIs. Accessed: 2024-10-06.

Micha Elsner and Kiwako Ito. 2017. An Automatically Aligned Corpus of Child-Directed Speech. In *Proc. Interspeech 2017*, pages 1736–1740.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Larry Fenson, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development*, pages i–185.

Frank Förster, Chrystopher L. Nehaniv, and Joe Saunders. 2011. Robots that say 'no'. In *Advances in Artificial Life. Darwin Meets von Neumann*, pages 158–166, Berlin, Heidelberg. Springer Berlin Heidelberg.

Frank Förster, Joe Saunders, Hagen Lehmann, and Chrystopher L. Nehaniv. 2019. Robots learning to say "no": Prohibition and rejective mechanisms in acquisition of linguistic negation. *J. Hum.-Robot Interact.*, 8(4).

Zhongqiang Huang, Lei Chen, and Mary Harper. 2006. An open source prosodic feature extraction tool. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jingbei Li, Yi Meng, Zhiyong Wu, Helen Meng, Qiao Tian, Yuping Wang, and Yuxuan Wang. 2022. Neufa: Neural network based end-to-end forced alignment with bidirectional attention mechanism. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8007–8011. IEEE.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Rochelle S. Newman, Meredith L. Rowe, and Nan Bernstein Ratner. 2016. Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, 43(5):1158–1173. Epub 2015 Aug 24.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Joe Saunders, Hagen Lehmann, Frank Förster, and Chrystopher L. Nehaniv. 2012. Robot acquisition of lexical meaning - moving towards the two-word stage. In *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–7.

Joe Saunders, Hagen Lehmann, Yo Sato, and Chrystopher L. Nehaniv. 2011. Towards using prosody to scaffold lexical meaning in robots. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–7.

Matthew Saxton. 2017. *Child language: Acquisition and Development*. Sage Publications Ltd, London, UK.

Emmanuel Senft, Séverin Lemaignan, Paul E. Baxter, Madeleine Bartlett, and Tony Belpaeme. 2019. Teaching robots social autonomy from in situ human guidance. *Science Robotics*, 4(35):eaat1186.

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.