



# A Hierarchical Predictive Processing Approach to Modelling Prosody

Juraj Šimko<sup>1</sup>, Adaeze Adigwe<sup>1,2</sup>, Antti Suni<sup>1</sup>, Martti Vainio<sup>1</sup>

<sup>1</sup>University of Helsinki, Finland

<sup>2</sup>Readspeaker, Netherlands

firstname.secondname@helsinki.fi

## Abstract

Prosodic patterns—and linguistic structures in general—are hierarchical in nature, providing for efficient means for encoding information in temporally constrained situations where communicative events occur. However, there are no theoretical frameworks that are capable of representing the full extent of linguistic behaviour in a cohesive way that could capture the paradigmatic and syntagmatic links between the organizational levels present in everyday speech.

Here we propose a novel theoretical and modelling account of perception and production of prosodic patterns in speech communication, derived from the influential Predictive Processing theory of neural implementation of perception and action based on a hierarchical system of generative models producing progressively more detailed probabilistic predictions of future events. The framework provides a conceptualization of the hierarchical organization of speech prosody as well as a principled way of unifying speech perception and production by postulating a single processing hierarchy shared by both modalities. We discuss the possible implications of the theory for prosodic analysis of speech communication, including conversational setting. In addition, we outline a viable computational implementation in the form of a machine learning architecture that can be used as a testbed for generating and evaluating predictions brought forth by the theory.

**Index Terms:** predictive processing, prosody perception, prosody production, conversation, machine learning architecture

## 1. Introduction

The role of prosody in encoding many aspects of speech communication, and providing a framework for speech acts is widely acknowledged. One of the most important insights of prosodic research is the *hierarchical* organization reflecting multilayered contextual aspects of speech communication. The nesting of units of prosodic analysis—syllables within (prosodic) words, words within phrases, phrases within utterances, utterances within speech acts—is conceptualized by, for example, the phonological metrical theories of speech prosody that assume that the prosodic hierarchies are governed by similar (albeit not identical) principles as the hierarchical syntactic structure of utterances (e.g., [1]). Some theoretical accounts explicitly embrace the hierarchies as an organizational principle for prosodic-acoustic characteristics of speech signal (e.g., [2, 3]). And while some other influential theoretical accounts downplay the impact of hierarchicity on prosodic patterns, they often implicitly acknowledge the hierarchical nature of influences behind formation of the patterns (intonational phrase and break hierarchies in ToBI [4, 5]; hierarchies of communicative functions in PENTA [6]).

Prosodic phenomena—durational and rhythmic patterns,  $f_0$  contours, intensity variation, changes in voice quality—are routinely analysed within clearly temporally delimited portions of speech signal (prosodic units). The patterns characterising a stressed syllable, for example, provide a paradigmatic contrast (stressed/unstressed) manifested by their syntagmatic relations to the other units (syllables) within a word (a higher prosodic unit). When the same word is focused, this paradigmatic distinction (with respect to its unfocused rendition) is brought forth, among other things, by adjustments in relational (syntagmatic) prosodic characteristics of its syllables. Going up the hierarchical ladder, the syntagmatic relationships marking the characteristics of the focused word within an utterance would reflect the intended higher-level contrast: the choice between the utterance with broad vs. narrow focus. And this choice itself (and its syntagmatic fit with the surrounding utterances) would be determined by yet wider contextual influences within the given discourse.

These influences span progressively wider and wider temporal scopes. The information (in a broad sense) encapsulated within “discourse units” or “topic units” has consequences for the choice of prosodic realization of lower-level discourse constituents, signalling, for example, the illocutionary force of utterances (statement, question, acknowledgement...); this choice is again reflected in the more detailed prosodic patterns lower down [7]. Even more global pragmatic aspects, such as the flow of emotions, affective meanings, provide background for the discourse structure itself. The prosodic manifestations of syntagmatic relationships among the units on one level can thus be seen as *conditioned* by paradigmatic contrasts at the level(s) above.

In order to maintain cohesion among the units at a given level (necessitated by both production and perception/parseability constraints), there needs to be a strong lateral link between the successive units: at the lowest level, the possible continuations of the  $f_0$  contour are to a large extent *dependent* on the previous stretch of the pitch movement; higher up, a question intonation (alongside the intonation patterns of the previous speech material) considerably limits the possibilities for intonational patterns of subsequent speech turn. This causal link can be conceptualized in terms of predictability of the future events from the past events at the same hierarchical level. We suggest—in line with the more general theory of Predictive Processing—that it is this generative, predictive process of unit sequencing that is conditioned, in a top-down manner, by the higher-level assumptions (contrastive choices) reflecting progressively longer contextual dependencies, including dialogue-wide global characteristics of the communicative situation.

Prosody is not limited to relatively short distance phenomena, traditionally delimited by units such as intonational phrases and utterances. It is instrumental in monitoring and maintain-

ing discourse structure (“discourse units”, “topic units”; [8]). In conversational settings, speech and its prosodic patterns are simultaneously produced and perceived, and used—along with other devices such as syntax and lexicon—to help align the “belief states” of the interlocutors with each other and with the general flow of the conversation. Conceptualizing these belief states as the high-level assumptions from the previous paragraph, the interlocutors can be seen as striving to “reverse engineer” these high-level conditioning influences from the perceived stream of lower-level acoustic-prosodic events. On the one hand, correctly inferred conditioning influences will assist the dialogue partners with micro-management of some aspects of conversation dynamics, for example, predicting the current speaker’s intention of yielding or holding up to the turn, recognizing a topic change, etc. On the other hand, when contributing to the conversation, the speaker can use these re-constructed influences to do so in a coherent and natural way, in line with the overall flow of the dialogue, and, in turn, use prosody to influence the high-level representations and “belief states” of the dialogue partner.

In the following section we present a brief overview of the cognitive Predictive Processing framework that serves as a basis of our proposed account of perception and production of speech prosody introduced in Section 3. Subsequently we describe an extension of the approach to a conversational setting and its possible computational implementation.

## 2. Predictive Processing framework

Several investigations and theoretical treatments have highlighted various roles of predictions and predictability in speech communication and prosody (e.g., [9, 10, 11, 12, 13]). What we propose, however, is considerably bolder: *all* aspects of prosody, its production, perception, acquisition, in *all* scenarios (read speech, monologues, conversations, arguments...) can be fruitfully conceptualized in terms of a hierarchical predictive processing framework.

The framework we refer to in this proposal is the influential and highly developed Predictive Processing theory of neural implementation of perception and action in general (PP; [14, 15, 16, 17, 18]). This theory assumes a system of generative models operating at various time-scales. These *hierarchically organized* models generate progressively more and more detailed probabilistic predictions of future events in a “top-down” and lateral manner. That means that each model generates its outputs based on the present states (potentially also encoding the memory of the past) and on the outputs from the models at higher hierarchical levels. The outputs of each model are predictions, in the form of statistical distributions over the future possible states.

The working system ensures that the “conditioning” inputs generated at higher levels minimize the prediction error at the levels below; the lowest level prediction error directly capturing the mismatch between the low-level predictions and low-level percepts. The proponents of the Predictive Processing hypothesis suggest that this congruence is achieved through Bayesian inference: the probabilistic input from the layer above is treated as a prior, and the actual value of predicted event as evidence used to calculate posterior probabilities. When the inference process is completed (in a “bottom up” way), the models higher up the hierarchy generate more and more abstract hypotheses of “what’s going on”, all geared towards minimizing the prediction errors, in particular the difference between the predicted immediate percepts and the percepts themselves.

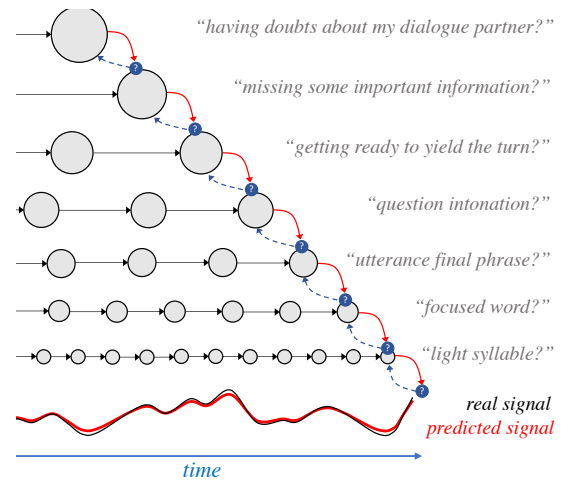


Figure 1: A schematic sketch of the proposed architecture. The hierarchy of the generative models building upon the past states is depicted in the left. The progressively longer horizontal arrows indicate longer temporal context shaping the predictions (red arrows) of higher levels. The blue dotted arrows symbolise the influence of bottom-up error. On the right hand side we list a hypothetical prosodic and discourse interpretation of the prediction at various levels.

This sketch of the Predictive Processing framework (for a more thorough description see, e.g., [18]) is primarily about perception, and this is despite the fact, that at its core is an interacting system of *generative* models. The theory, however, aspires to provide an explanation for action as generated by systems employing this cognitive strategy. In essence, the action is the response of the system that for some reasons opts to react to a mismatch between predictions and evidence by adjusting the *evidence* (i.e., its low level percepts) rather than it’s hierarchical system of beliefs (imagine, for example, adjusting the viewing angle when a glare on a window covers expected objects within your visual field, rather than accepting the belief that these objects simply vanished).

The framework thus principally combines perception and action. To a large extent, both modalities make use of the *same* stack of generative models. In perception, the system strives to find a hierarchy of progressively more and more abstract states that provide the best overall explanation (capture the sense of) for what’s going on in the world as presented to it in the form of its direct percepts. When acting, the same hierarchy is behind producing the most appropriate behaviour in the given global context.

Learning in this paradigm is equivalent to creating and fine-tuning the rich hierarchy of appropriate hypothesis generating models, including lateral and conditioning connections, that are together capable of predicting any patterns relevant for the task (or a set of tasks) in a way that enables the agent to appropriately and promptly act in its environment.

## 3. Predictive Processing in Prosody

As outlined in the introduction, speech prosody with its hierarchical organization naturally yields itself to this hierarchical predictive account. Figure 1 presents a sketch of the proposed architecture. Several generative models at the bottom of the hierarchy can be seen as directly corresponding to the prosodi-

cally relevant hierarchical levels found in speech. At the lowest level, where the prediction meets with actual low-level auditory information, the prediction of the immediate future in terms of prosodic signals ( $f_0$ , intensity, voice quality...) is made based on their past history and conditioned by predicted prominence realization of the constituent syllable (stressed/unstressed). The syllable characteristics are in turn predicted based on the characteristics of the preceding syllables and conditioned by features pertaining to a higher metrical level, e.g., that of prosodic word. And the characteristics of the prosodic word (focused? pitch-accented?) are conditioned by the nature of the phrase, which is determined by the prosodic characteristics of the utterance, etc.

At some point, the match between the recognized prosodic metrical units and the necessary systems of predictive models becomes less tangible. We nevertheless hypothesize the existence of models predicting characteristics with progressively wider scopes that correspond to the flow of more and more abstract features of spoken interaction: conversational turns, discourse topics, overall mood of discourse.

It is important to acknowledge, that while prosody plays a role in determining and signalling many local and global features of discourse and interaction, it does it as a part of a considerably more complex linguistic machinery. This machinery includes structural constraints conceptualized as grammar, syntax, semantics. Importantly, these aspects of speech communication also manifest hierarchical organization, with units such as syllables, words, phrases actually considerably overlapping with the prosodic hierarchy. In fact, the hierarchical predictive processing account presented in Section 2 could be adapted to syntax and semantics (work in this direction has been reported in [19, 20]). The syntactic and semantics predictive processing systems can be thus seen as interacting with the proposed prosodic one.

Even more interestingly, these systems can be seen as unified in a single predictive processing architecture, with different aspects interacting and informing each other. Computational implementations of such a unified system (see Section 5) can potentially be used to disentangle the complementarity of different communicative means (prosody, syntax, semantics) including the known discrepancies between syntactical and prosodic hierarchies.

#### 4. Prosodic Predictive Processing in Conversations

The *listener's* prosodic predictive processing architecture compares the prosodically relevant aspects of the speech signal produced by the *speaker* (such as  $f_0$ , intensity, spectral slope; presumably extracted by early auditory processing mechanisms) to predictions of the lowest level in the hierarchy. As long as the two match, nothing much happens (one can argue, that in that case there is no actual percept for the listener in terms of prosody). When there is a discrepancy—the actual prosodic signal is deemed unlikely according to the presently favoured hierarchy of the listener's hypotheses—this information ('error signal') is used to adjust the hypotheses in a bottom-up manner. These adjustments continue until a new hierarchy of hypotheses is found, generating correct predictions. When this is achieved, the listener, in essence, *gets it*; her "belief state" is aligned with—even though not identical to—that of the speaker.

According to the Predictive Processing proposal, this hierarchy of hypotheses drives a hierarchy of generative models.

The listener can directly use this hierarchy to govern her own production of the speech signal, by, for example, initiating her own turn in the conversation. This, of course, does not mean that her turn will consist of a straightforward continuation of what the conversation partner (the original speaker) would have said had he not yielded the turn (although this also can, and does, happen). Mastering the conversation skills means figuring out what to say (and how) in order to contribute to co-creating the aligned set of hierarchical models. The listener can in fact also "bargue in" when the aligning of her predictions with what's going on would require too great, perhaps unacceptable, adjustment in her set of high-level hypotheses. In this case, her intent will be to nudge the generative hierarchy of the conversation partner in a reciprocal fashion. ("Belief state" alignment is not—should not!—be a one-way process.)

The production requires its own set of generative predictive models responsible for generating a sequence of appropriate neural signals that, in case of prosody, primarily drive the behaviour of laryngeal and respiratory apparatuses. This production hierarchy could be attached at the bottom of the overall predictive hierarchy and conditioned by the set of higher level models. (This production system can be seen as a fork at the bottom of the hierarchy parallel to a low-level perception fork in charge of processing the auditory consequences of the physical signal).

This architectural proposal has two important theoretical consequences. First, perception and production of prosody are fully integrated. Rather than treating these two modalities as parts of a perception-production loop with two distinct pathways meeting high up and requiring an additional central processing apparatus, they are conceptualised as two facets enabled by a fully (or almost fully) shared predictive processing architecture.

Second, conversation is not a ping-pong of messages generated by one interlocutor, *subsequently* processed by the other one who in turn comes up with her own message, and passes it on to be processed by the waiting conversation partner. Rather, it is a fully parallelized, shared activity, with ongoing adjustments and aligning of the "belief states" of the interlocutors, and co-creating the overall conversational content and context. This sense-making activity can be conceptualized in an enactivist language of co-creating shared meaning through dynamical coupling between the conversational partners' own prosodic and linguistic dynamics.

#### 5. Speech technology architecture

A technologically minded reader can see the proposed model of hierarchical predictive processing of prosody as a blueprint for a technological implementation. While we are at this point not aware of a full artificial neural network implementation of the Predictive Processing architecture, the recent development in the field of machine learning provides important building blocks that can be used to create such a system.

For example, the recently proposed Contrastive Predictive Coding (CPC) architecture [21] and its variants (e.g., [22]) is an unsupervised learning approach that derives higher-level latent representations from a signal, capturing temporally less fine-grained regularities used to predict the future signal samples. The approach takes advantage of a probabilistic contrastive loss which induces the latent space to capture information that is maximally useful for generating predictions. The technique has been successfully implemented in a speech recognition task, where the latent higher-level representations encode informa-

tion (with wider temporal scope compared to the signal sampling rate) interpretable in terms of differences among regularly occurring patterns, e.g., speech segments [21].

The CPC approach differs from the presented account of Predictive Processing in many important aspects. First, instead of the Bayesian inference approach it uses the principle of maximizing the mutual information between the encoded representations. Second, it is not hierarchical; rather the architecture is limited to a single-level abstraction roughly corresponding to one layer in the Predictive Processing hierarchy.

Nevertheless, we believe that CPC can be used as a baseline implementation of vanilla predictive coding. The next natural step would then be to implement a hierarchical architecture with a vertical stack of CPC models with latent representations with progressively wider and wider temporal receptive fields (portions of the signal that influence their values), reminiscent of the WaveNet speech synthesis architecture [23].

The Predictive Processing implementation using these building blocks can be trained on appropriate speech material containing longer stretches of spoken interaction. Our expectation is that the higher hierarchical layers will learn (in a largely unsupervised manner) progressively more and more abstract latent representations capturing wider and wider contextual information relevant for predicting future signal samples. In the production (speech synthesis) modality, these representations can then be used to condition the generation of the appropriate, prosodically rich, and context aware speech signal.

In a monologue (non-dialogue data) setting, the generative character of the predictive system will yield context aware renditions of prosodic patterns. When trained on a conversational data, the system will, it is our hope, learn how to extract the long-term contextual prosodic characteristics from speech of one conversational partner and use them to condition the generation of prosodically appropriate dialogue contributions.

If the architecture built along these sketchy guidelines works, it will undoubtedly provide an important contribution to speech technology. Equally importantly, the implementation—even a partial one—will serve as a *computational model* of the theoretical framework proposed here. This model can be used as a developing testbed for generating replicable predictions for rigorous and transparent evaluation of the proposed approach.

## 6. Discussion and further questions

While these connectivist computational models will be trained on suitable corpora of conversational speech (i.e., in a supervised manner), the predictive hierarchies themselves are envisaged to be formed in an essentially unsupervised way (with no explicit “ground truth” hierarchical signal to govern their formation). The hierarchies and lateral links are thus assumed to arise in a way that best meets the requirements of prosodic cohesion and context dependence of prosodic patterns, and encoded, in machine learning terms, in the form of latent representations. This means that there is no guarantee that the hierarchical structures captured by the computational implementation of the Predictive Processing architecture will directly, “veridically” correspond to the assumed units of the prosodic hierarchy—syllables, words, phrases, and so on—that inspired the present account in the first place.

Our hypothesis schematically captured in Fig. 1 and to be tested by the computational models, is that the latent conditioning predictions of the several lowest hierarchical levels *will* correspond to the paradigmatic contrasts as traditionally conceptualised in prosodic analysis. That is, the condition-

ing signals generated by these lower-level models will correlate with phonologically meaningful characterizations of contrasts such as, respectively, stressed/unstressed syllable, non-accented/accented/focused word, etc.

For the higher-level predictive models, however, this type of phonological interpretation will be progressively more and more tenuous. We hypothesise that these higher level of hierarchy, capturing wider contextual information, will be best interpretable in terms of global discourse structure informational and pragmatic characteristics. When combined with hierarchical Predictive Processing implementation of other aspects of linguistic communication as proposed above (e.g., predictive inferential systems capturing syntactical and semantic structures, cf. [19, 20]), the system has a potential, so is our belief, to account for complex aspect of human speech mediated communication, including rudimentary understanding. Moreover, this envisaged Predictive Processing architecture can be used to investigate the direct role that prosody plays in facilitating this communication and its aspects (e.g. by comparing models with and without prosodic processing).

As suggested above, the proposed framework combining perception and production of speech prosody (and, in extension, of speech communication in general) is best evaluated in conversational setting, on conversational material. The proposed approach might shed light on several research questions in this research area. For example, the acoustic-prosodic entrainment, where the dialogue partners align in some acoustic and prosodic aspect of their speech (e.g., [24]), may be a direct consequence of perceptual mechanism being driven by predicting the interlocutors acoustic and prosodic patterns. The similarity might simply result from a “leak” of this predictive perception system to the production system sharing the generative hierarchy.

On a more conceptual level, Predictive Processing provides a natural framework for probabilistic phenomena such as turn-taking. It also implies the need of hierarchical analysis in order to generate more robust theoretical and practical predictions of, e.g., turn-taking behaviour involving discourse-wide contextual influences rather than simple cues derived from purely low-level patterns.

The general Predictive Processing framework is a developing paradigm, incorporating many cognitive phenomena. For example, it strives to provide a principled account for attention, with an attention mechanism conceptualized in terms of error-weighting mechanism. Mismatch errors at various hierarchical levels are selectively weighted up or down depending on what level of attention is paid to the corresponding predicted patterns; this sensitivity is, in turn, tuned using the other predictions generated by the model (see [18] for details).

Re-conceptualization of this and other insights of the Predictive Processing framework in terms of speech prosody and linguistic processing will open up new avenues for studying and modelling phenomena like using prosody (prominence) for managing interlocutors’ attention, links between predictability, production and perception of prosodic units (e.g., [9, 11]), and other manifestations of speech prosody.

## 7. Acknowledgements

The first and third authors’ contribution is partially funded by the Academy of Finland Fellowship. The second author has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 859588.

## 8. References

- [1] B. Hayes, "The prosodic hierarchy in meter," in *Rhythm and meter*. Elsevier, 1989, pp. 201–260.
- [2] E. Saltzman, H. Nam, J. Krivokapic, and L. Goldstein, "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation," in *Proceedings of the 4th International Conference on Speech Prosody (Speech Prosody 2008), Campinas, Brazil*. Citeseer, 2008, pp. 175–184.
- [3] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.
- [4] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling english prosody," in *ICSLP*, vol. 2, 1992, pp. 867–870.
- [5] J. Pierrehumbert and J. B. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgan, and M. Pollack, Eds. MIT Press, 1990, pp. 271–311.
- [6] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Communication*, vol. 46, no. 3-4, pp. 220–251, 2005.
- [7] M. E. Beckman and J. B. Pierrehumbert, "Intonational structure in japanese and english," *Phonology*, vol. 3, pp. 255–309, 1986.
- [8] A. Gravano, J. Hirschberg, and Š. Beňuš, "Affirmative cue words in task-oriented dialogue," *Computational Linguistics*, vol. 38, no. 1, pp. 1–39, 2011.
- [9] M. Aylett and A. Turk, "The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, vol. 47, no. 1, pp. 31–56, 2004.
- [10] Y. Hirose and R. Mazuka, "Predictive processing of novel compounds: Evidence from Japanese," *Cognition*, vol. 136, pp. 350–358, 2015.
- [11] S. Kakouros and O. Räsänen, "Perception of sentence stress in speech correlates with the temporal unpredictability of prosodic features," *Cognitive science*, vol. 40, no. 7, pp. 1739–1774, 2016.
- [12] N. Henry, H. Hopp, and C. N. Jackson, "Cue additivity and adaptivity in predictive processing," *Language, Cognition and Neuroscience*, vol. 32, no. 10, pp. 1229–1249, 2017.
- [13] N. Henry, C. N. Jackson, and H. Hopp, "Cue coalitions and additivity in predictive processing: The interaction between case and prosody in L2 German," *Second Language Research*, p. 0267658320963151, 2020.
- [14] K. Friston, "Learning and inference in the brain," *Neural Networks*, vol. 16, no. 9, pp. 1325–1352, 2003.
- [15] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.
- [16] K. Friston, "Hierarchical models in the brain," *PLoS computational biology*, vol. 4, no. 11, p. e1000211, 2008.
- [17] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behavioral and brain sciences*, vol. 36, no. 3, pp. 181–204, 2013.
- [18] —, *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press, 2015.
- [19] A. Prefors, T. Regier, and J. B. Tenenbaum, "Poverty of the stimulus? A rational approach," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 28, no. 28, 2006.
- [20] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.
- [21] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [22] J. Chorowski, G. Ciesielski, J. Dzikowski, A. Łancucki, R. Marxer, M. Opala, P. Pusz, P. Rychlikowski, and M. Stypułkowski, "Aligned contrastive predictive coding," *arXiv preprint arXiv:2104.11946*, 2021.
- [23] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
- [24] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.