Tools for Estimating the Perceived Level of Phonetic Reduction

Nigel G. Ward[1,2], **Javier Vazquez**[1], **Emma R. Boushka**[1], **Oliver Niebuhr**[2]
[1]University of Texas at El Paso, El Paso, Texas, USA
[2]University of Southern Denmark, Sønderborg, Denmark
nigelward@acm.org, jvazquez073015@gmail.com, erboushka@miners.utep.edu, olni@sdu.dk

Abstract

Phonetic reduction is very common in casual speech, where it is associated with several important pragmatic functions. However these phenomena have been little studied. To support investigations and applications, we present tools that automatically estimate the level of perceived phonetic reduction. Trained on annotated dialog data and exploiting HuBert features, these handle American English and Northern Mexican Spanish. For English, word-level predictions correlate up to 0.55 with average human judgments. This is adequate at least for statistical studies of reduction in corpora, as seen in explorations of turn-yielding and prominence-marking behaviors. The tools are open-source and publicly available.

Keywords: reduced articulatory precision, hypoarticulation, degree of articulation, American English, Northern Mexican Spanish

1. Introduction

In human speech, phonetic reduction is very common, especially in casual speech. Also referred to as hypoarticulation or reduced articulatory precision, reduction is a complex phenomenon (Ernestus and Warner, 2011; Jaeger and Buz, 2017; Cangemi and Niebuhr, 2018; Zellers et al., 2018; Niebuhr, 2017).

Proper application of reduction is important in speech synthesis: speech with every phoneme enunciated is unnatural. While early research strove to elucidate the factors involved — such as predictabilty, prosodic contexts, and local syntactic structure — modern synthesizers learn to control reduction implicitly. While they do very well in terms of naturalness, at least for read-style speech, this is not sufficient for all use cases. The growing demand for speech synthesizers to be more useful in more scenarios implies that they will need expanded communicative range, and, in particular, the ability to effectively convey more pragmatic functions (Wagner et al., 2019; Marge et al., 2022; Mohan et al., 2021; Lameris et al., 2023).

Reduction has great potential value here. Although its functions remain poorly understood, it is involved in, at least, turn hold in English and German (Local and Walker, 2012; Niebuhr et al., 2013), repair and agreement in Dutch (Plug, 2005, 2011), sarcasm in German (Niebuhr, 2014), and, in English, self-deprecating asides (Gustafson et al., 2023b), cues for backchannel feedback, expressions of a self-sufficient stance, topic closings (Ward, 2019), and positive assessment (Ward et al., 2025a). Indeed, it has been argued that phonetic reduction can pattern with other prosodic features and should be included in that set, along with voicing properties and the traditional, pitch, energy, and duration (Beller et al., 2008; Niebuhr, 2016; Ward et al., 2025a). Better control of reduction in speech synthesis is possible (Picart et al., 2014; Gustafson et al., 2023a), so the time is ripe to prepare to exploit it in practice.

Progress has been impeded, however, by the lack of a suitable tool for automatically detecting reduction in speech. Such a tool could have value in at least two ways. First, it could enable loss functions for synthesizer training to include a penalty for inappropriate use of reduction or inappropriate lack thereof, as a step beyond pitch-only prosody loss functions on the road to more complete and accurate ones (Huang et al., 2023; Ward et al., 2025b). Second, such a tool could support basic linguistic studies to identify all the roles of reduction in conveying pragmatic functions, especially through corpus studies. These advances could, ultimately, enable dialog systems that are better able to recognize speaker intents, and also able to produce utterances that more clearly convey various pragmatic functions.

2. Problem Statement

We aim to build tools with

Input:	a speech signal, typically a conversational utterance, and
Output:	an estimate of the perceived degree of reduction across the utterance, either frame-by-frame or word-by-word.

This framing of the problem, inspired by (Ward and Ortega, 2024), was chosen as the most relevant for the goals noted above: better speech synthesis and pragmatic function investigations. In particular, we aim to model the *percept* of reduction,

as it is likely more directly relevant to communicative functions than is any acoustic correlate.

3. Previous Approaches

Previous work on the automatic detection of reduction can be classified into correlate-based approaches, intelligibility-based approaches, and perception-based approaches.

3.1. Correlate-Based Approaches

In linguistic studies of reduction, most attention has been directed at various acoustic correlates, which reflect both general articulatory shortcuts and the application of language-specific phonological rules. For (semi-) automatic detection of reduction, word duration has been commonly used. While shortening is indeed a proxy for reduction (Jurafsky et al., 1998), and it can be easily detected from transcripts, when they exist, shortening is only one form of reduction (Burdin and Clopper, 2015), and the correspondence to the general percept is unknown.

3.2. Intelligibility-Based Approaches

Another body of work, more applications-oriented, has aimed to characterize the overall level of reduction of a speaker, as a measure of intelligibilty.

For learners, a common approach is to use Goodness of Pronunciation or similar estimates based on where a speech recognizer fails or has low confidence (Tu et al., 2018; Lubold et al., 2019; Arias-Vergara et al., 2023). These techniques presuppose the existence of transcribed data or high-quality speech recognizers, but these do not exist for all speech genres and all languages. Other work has applied deep learning to the detection of reduction in learners' speech (Chen et al., 2022), although this was only evaluated on clear cases.

For medical needs, for example to evaluate whether a surgical intervention was beneficial, "articulation entropy" as an acoustic measure and measurements of articulator displacement using electromagnetic articulography both have promise (Lee et al., 2006; Jiao et al., 2016). As these methods have been tested only for making estimates from relatively long samples of speech, they are not yet suitable for use at the time scale at which reduction contributes to the expression of pragmatic functions (Niebuhr et al., 2013; Niebuhr, 2014, 2016; Ward et al., 2025a), namely reduction at the time scale of words.

Here, rather than consider reduction to be a mere indication of poor intelligibility (Quintas et al., 2023), we aim to model it in its own right. While it is true that people often perceive speech as more reduced when it is more difficult to understand, or departs from the norms of careful speech, it is also true that many instances of reduction are perfectly intelligible in context. As a consequence, the degree of reduction is a "speech feature" that is potentially available to speakers and languages as a resource for conveying meaning, and indeed often seems to serve as such, as noted above.

3.3. A Perception-Based Approach

For these reasons, we chose to model *perceived* reduction. This, however, brings challenges: the lack of a standard definition and the variation in what people perceive as reduced, as noted in the only previous attempt to model perceived reduction (Ward and Ortega, 2024), where low inter-annotator agreement was flagged as a concern. Beyond some observations in that paper, the causes and extent of such perceptual differences do not seem to have been examined, and we feel that this should be a priority for future research. However, people can still communicate despite these differences, so incomplete knowledge here need not delay the practical application of reduction modeling.

The only previous work that tried to model perceptions of reduction was our own (Ward and Ortega, 2024), which used a very similar problem definition, also chosing to focus on dialog data, rather than read speech or monologue, and to treat reduction as a gradient phenomenon, rather than as simply present or absent. The approach was to use prosodic features, notably pitch, spectral tilt, and harmonicity, as predictive features, inspired by various reports of correlations between these features and reduction or hyperarticulation. That model, however, performed poorly, correlating only (0.243) with human judgments.

Thus, previous work has not met the need for a generally usable method for estimating the degree of reduction in speech.

4. Data

For our primary datasets we used previously released resources (Ward et al., 2025a): these were used for all training. For evaluation we used those same data sets in leave-one-out fashion, plus three new sets.

The training data is described in full in (Ward et al., 2025a), but it is worth here noting some key properties: It consists of pragmatically-rich conversations labeled for degree of reduction at the region level. Most regions being a single word, below we refer to these also as word-level annotations. There are 31 stereo minutes of English conversations and 25 of Spanish conversations. The sets for the two languages are matched in two ways: all the speakers came from the same demographic — namely Spanish-English bilingual college students living in

greater El Paso, Texas — and the annotations were all done by one individual, C, from the same demographic and phonetically naive. The rubric was "subjective judgments of being poorly articulated or possibly hard to understand without context," according to which each region was labeled 0 for enunciated, 1 for normal, 2 for reduced, or 3 for highly reduced. Regions with only non-lexical utterances (*uh-huh*, *umm*, etc.) or laughter were not annotated. Both tracks of each conversation were annotated. Of the 3051 annotated regions for English, 27% were either reduced or highly reduced, and for Spanish, 31% of the 1614 regions.

We supplemented this data with three additional sets of annotations, done following the same protocol. We accepted that perceptions of reduction differ, and chose to respect this, rather than try to rigidize the protocol to improve conformity.

The first new data set was designed to support creation of a gold standard and to enable examination of inter-annotator agreement. For this, all three authors independently annotated all regions in 10 minutes from two of the English dialogs of the primary data, chosen for the high engagement level of the participants.

The other two data sets were added to examine generality across corpora, including variation in recording conditions, genre, and demographics. First, one of us, D, annotated 4 stereo minutes from a corpus of telephone conversations (Godfrey et al., 1992)(Godfrey et al., 1992). Unlike our primary data, this was a conversation between strangers, who were middle-aged and had an East Texas accent. Second, the same author annotated 9 minutes of one speaker in a dialog from a task-oriented dialog corpus (Lehnert-LeHouillier et al., 2020). This speaker was an adolescent interacting with a graduate student, recorded through a desktop microphone. Unlike the primary data, in this last data set the reduction levels were bimodal, with 12%, 43%, 11% and 35% of the frames being at levels 0, 1, 2, and 3, respectively. All these annotations are released at `www.cs.utep.edu/nigel/reduction/` .

5. Models

Our basic strategy is to leverage the power of embeddings in speech pretrained models, that is, models trained by self-supervised learning. The internal features of these models have been shown to represent information that is useful for many similar discriminations and tasks (Lin et al., 2022; English et al., 2023; Chernyak et al., 2022). Here we investigate, for the first time, whether these also contain information relevant to reduction. Following common practice, we use pretrained-model features as input to a simple downstream model which we train for our specific task.

Specifically, we chose the HuBERT Base model (Hsu et al., 2021), for two reasons. First it performed well for other prosody-related tasks (Lin et al., 2022). Second its features have been shown to work well for emotion recognition, speaker identification and speech recognition, so they likely represent both low-level acoustic detail on the phonetic realization and higher-level information, including the speaker's intended phoneme sequence.

Given an audio file, this outputs 12 layers of 768 features for each 20 millisecond frame, for each track. Here we chose to use only the 12th-layer features, which are generally good for many tasks (Lin et al., 2022; Yang et al., 2024), and slightly outperformed the other layers in pilot studies.

Our first task was frame-level predictions, for which the prediction target was the label inherited from the annotated region encompassing the frame. For this the 768 features were fed into a simple linear regression model. We chose linear regression to minimize the number of free parameters, as a bid to avoid overfitting, given the limited size of the data, and because linear decision heads can perform very well for tasks like this (Narain et al., 2025). We evaluated models in leave-one-dialog-out style.

The second task was word-level predictions. For this we tried two methods. The first method averaged features for all frames within the labeled region, and then fed the vector of feature averages to the linear regression model from the first task, but here to predict the word-level reduction. The second method used the average of all the frame-level predictions within the region.

The time and memory requirements here are dominated by the first step: running HuBert to compute the features. This takes less than one second per second of stereo audio on a commodity laptop, but requires about 2GB per stereo minute to process each file. We note that we compute these features by file rather than by utterance, as margins of a couple of seconds around any clip of interest are needed to ensure stable computation of the HuBert features.

In addition to this basic model, we tried variant models of four types: using feature downselection before linear regression, using subsets of the top PCA-derived features in an attempt to abstract away from various types of noise, using HuBert Large, and using an ensemble of per-speaker models.

6. Performance Metric and Results

We chose to measure performance by the correlation between the predictions and the human annotations. Correlations are appropriate because, for purposes of pragmatic meanings, it is more useful

Task		Correlation
Frame Predictions:		0.245
Word-Level Predictions:	by average of features	0.272
	by average of predictions	0.376

Table 1: Correlations between predicted and annotated values.

Model	Correlation
basic	0.376
downselected to 96 features	0.313
downselected to 165 features	0.331
PCA features explaining 50% of the variance	0.224
PCA features explaining 90% of the variance	0.326
PCA features explaining 99% of the variance	0.322
PCA features explaining 99.9% of the variance	0.361
HuBert Large	0.266
ensemble of per-speaker models	0.384

Table 2: Word-level correlations for different models.

to know whether a word is reduced relative to the norm for the current speaker than to know its absolute level of reduction. While correlations somewhat understate the true agreement, since the human ratings were elicited using discrete scales, they are adequate for making comparisons.

6.1. English

Table 1 shows the results for the primary data. While we see only modest correlations, even the weakest is statistically significant ($p < 0.001$, via the t statistic).

From the table we also see that predicting reduction at the frame level is harder. This was confirmed by plotting these predictions over time: they were very unstable and noisy. But, of course, predicting reduction at 20 ms granularity is not a real task: it is useful only for supporting predictions at wider time scales.

The table also shows that the second method for word-level prediction performs better, unsurprisingly, as it retains more information than the feature-averaging method. We also find that this model outperforms that of previous work (Ward and Ortega, 2024): 0.38 versus 0.24. This indicates the value of using HuBert features rather than a small set of prosodic features.

Table 2 shows the results with the variant models, all making word-level predictions using the second method: averaging frame-level predictions. None significantly outperforms the basic model. HuBert Large significantly underperforms, perhaps because its 1024 features make overfitting easy.

genre	an.	frame	word
face-to-face conversation	C	0.24	0.38
face-to-face conversation	D	0.23	0.29
telephone conversation	D	0.18	0.29
task-oriented dialog	D	0.22	0.43

Table 3: Correlations for different datasets. (Row 1: leave-one-dialog out, 6-fold over 31 minutes; Row 2: trained on 26 minutes, tested on 5 by a different annotator; Row 3 trained on 31, tested cross-domain on 4 minutes; Row 4: trained on 31, tested cross-domain on 9 mono minutes)

Table 3 shows that the model performs reasonably well also for other genres of English.

Table 4 shows the agreement both between annotators and between the model and each annotator, over the 10 minutes for which we have multiple annotations. We observe significant variation in how much the pairs of annotators agreed, although overall the level of agreement does not seem to be worse than seen previously (Ward and Ortega, 2024). We think that this level of agreement is not a flaw of our particular data, but an issue of essential variation among people in how they perceive reduction. In Table 4 we also see variation in how well the model matched the different annotators. While overall the model is performing below inter-annotator agreement, 0.43 versus 0.47, the difference is small, and the model sometimes outperforms some annotators: for example, it does better than annotator N as a predictor of J's percep-

	C	D	N	J	average
C	—	0.41	0.47	0.47	
D	0.41	—	0.47	0.58	
N	0.47	0.47	—	0.44	
J	0.47	0.58	0.44	—	
pairwise average	0.45	0.47	0.46	0.50	$\rightarrow$ 0.47
model	0.46	0.43	0.35	0.49	$\rightarrow$ 0.43
discretized model	0.45	0.42	0.35	0.47	$\rightarrow$ 0.42

Table 4: Frame-level agreement, correlations. The first four lines show the pairwise agreement among annotators, the fifth shows the average agreement between the annotator named in the column head and the other three annotators, the sixth shows the agreement between the model and the annotator named in the column head, and the seventh similarly shows the performance for the discretized model.

		Correlation
Frame Predictions:		0.151
Word-Level Predictions:	by average of features	0.121
	by average of predictions	0.262
”	English-trained model ”	0.188

Table 5: Results for Spanish

tions: 0.49 versus 0.44.

As a follow-on, we also built a discretized model. The main model outputs continuous values, but the human annotators were limited to integers. While this is a real advantage of the model, it makes the comparisons somewhat unfair. To overcome this, we created a variant model by simply discretizing the basic model's predictions, using thresholds chosen according to the distribution of all annotators' labels across the 10 minutes. As seen in the last line of Table 4, the correlations for this model were only slightly lower, and still mostly in the range of human performance.

Finally, for the 10 minutes labeled by all four annotators, we found the model's correlation with the *average* of the labels to be 0.55. As this average is the nearest thing we have to a gold standard, we report this as our headline performance number.

6.2. Spanish

Following the same approach, we also built a model for Spanish. As seen in Table 5, the results were lower than for English. Reasons for this may include: 1) the lower proportion of highly-reduced regions in Spanish, which makes the prediction task harder, and 2) the fact that the HuBert model used was trained only on English data. In addition, as a first stab at cross-language generality, we measured the performance of the English-trained model for predicting Spanish reduction; it was lower, unsurprisingly, given the differences in the phonotactics and typical reduction patterns of the two languages, such as the comparative paucity of consonant clusters and of vowel reduction in Spanish.

7. Qualitative Analysis

To better understand the performance and limitations of the basic model, we listened to the audio input at places where its predictions were furthest off, both overestimates and underestimates. We did both for the frame-level and word-level predictions. We did this for English only. We noted down any salient properties at that frame or across that region, and found that these often reflected a few common patterns of failure.

The most important factor was speaker identity, with the model's prediction's correlations ranging from 0.15 to 0.51 per speaker. We also noted that the model is better at predicting increased and decreased reduction than at predicting its absolute levels. In other words, the scaling was often off: for some speakers it generally underpredicted and for others generally overpredicted. This relates to an annotation tendency: the primary annotator seemed to consider reduction as relative to the norm for each speaker, which varied. If there are applications where better approximating this annotation style is important, speaker normalization would be appropriate.

Other factors related to the phonetic, lexical, prosodic and pragmatic properties of the speech. At the frame level, reduction overestimates often

featured the phonemes /v/, /l/, and various vowels near /e/, and reduction underestimates often involved /θ/. At the word level, the model often underestimated reduction on short words that were reduced to one phoneme, such as *and* reduced to /n/ and *if* reduced to /f/. Some prosodic properties were common in mispredictions: many of the overestimates were for cases of lengthening, and many of the underestimates were in utterances with low volume and/or creaky voice. Creaky voice can of course reduce intelligibility (Cammenga, 2018), so it is not surprising that many creaky regions were annotated as reduced (Ward and Ortega, 2024), but the two phenomena are not equivalent acoustically nor, probably, in how they contribute to conveying pragmatic functions. Some dialog acts were relatively common in mispredictions: many of the overestimates were for speakers holding the floor, which may relate to lengthening, and many of the underestimates related to false starts, which may relate to sharp transitions due to abrupt changes in articulatory intent.

We also found occasional "errors" due to the inadvertent inclusion of non-speech regions in some of the test data; this is unsurprising, since the concept of phonetic reduction is not meaningful for non-lexical items or laughter, let alone silence, so any model given such inputs will produce non-meaningful output.

These observations can help potential users decide whether and where to use this model. They also suggest avenues for its improvement, and give insight into both language features and annotator tendencies that future work should address. Of course, many of these problems could be addressed alternatively by trying better models, beyond linear regression, and, of likely even greater value, by expanding and improving the quantity and quality of the training data.

8. Examples of Use

Given the modest correlations, the predictions of this model cannot be relied upon for any specific speech sample. However they can be useful for statistical work, where individual inaccuracies can balance out over enough data. This section provides illustrations of the use of this tool for exploration and discovery.

First, we used the model to examine the relationship between reduction and turn yields, a classic research question for which evidence has been scant (Local and Walker, 2012; Niebuhr et al., 2013), doubtless due to the effort required to hand-label reduction. Our model enabled us to estimate the average degree of reduction in the vicinity of 11012 prototypical turn ends in Switchboard (specifically, timepoints at the end of at least 2 seconds of speech and followed within –0.25 to +0.75 seconds by the start of a turn in the other track lasting at least 2 seconds). Figure 1 clearly shows decreased reduction about 1 second before these turns end. This is compatible, as a contrapositive, with previous observations that reduction in English can be common at turn-hold points. However the graph also suggests that the picture is more complex, with other tendencies at other temporal offsets. In any case, the small magnitude of these effects is probably due to the fact that reduction is also affected by many other factors.

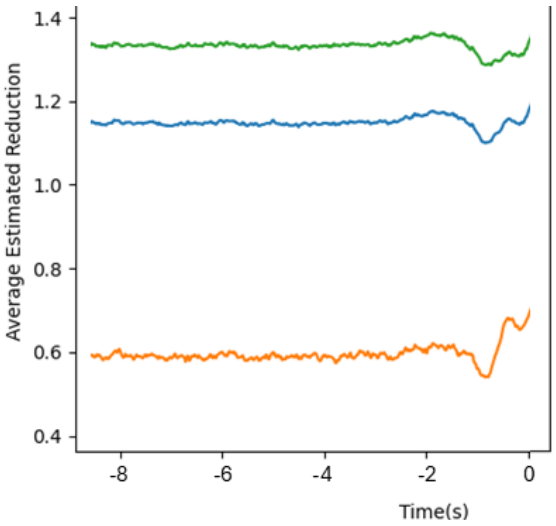

Figure 1: Average reduction level before turn yields, blue line, together with the 25th and 75th percentile averages, orange and green, respectively.

Second, we used the model to explore how reduction patterns with other prosodic features. While reduction has been argued to be part of the inventory of prosodic features, automatically processing it together with other features was never before possible. To do so, we added reduction as a feature along with other prosodic features, each computed over windows spanning three-second samples of speech from dialog. We then applyed principal component analysis to these features, a method shown to be able to reveal interesting prosodic patterns (Ward, 2019). Applying this to data from the Social Speech corpus (Ward et al., 2013), and examining the resulting principal components, we found reduction involved in several. Most interesting was one which captured a pattern in which reduction tended to last for a couple seconds, overlapping a longer region of high pitch, and preceeding a short region of louder, lengthened, and high-CPPS speech. Upon listening to some examples, it was clear that these were generally instances of emphasized words, preceded by a couple seconds of reduction. Thus this suggests a pattern in which reduction is present before emphasis.

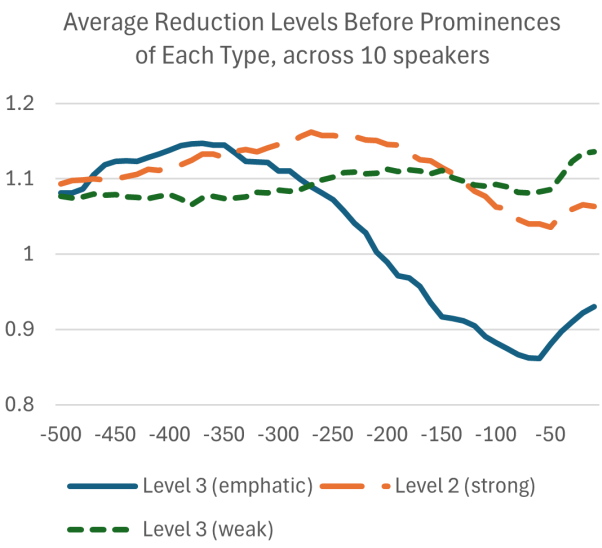

Figure 2: Average reduction level across 10 speakers before prominence peaks: emphatic, strong, and weak

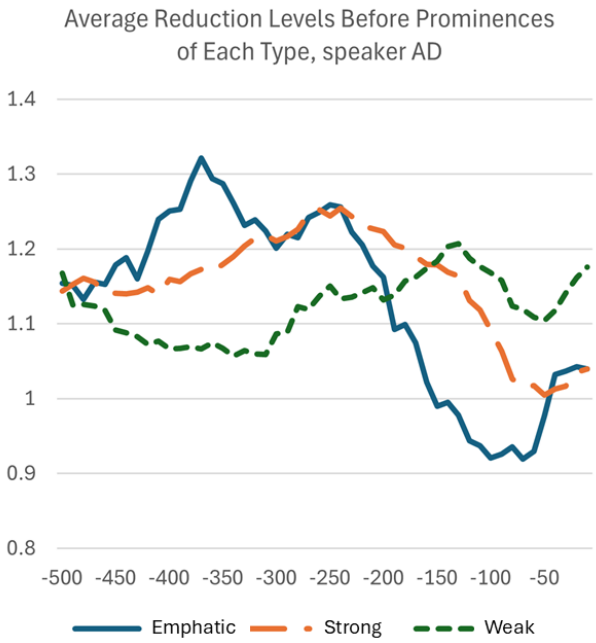

Figure 3: Average reduction level for Speaker AD before prominence peaks: emphatic, strong, and weak

This came as a surprise. The question of how emphasis and prominence are realized has been well-studied in prosody research (Kochanski et al., 2005; Watson, 2010; Andreeva et al., 2014; Cole et al., 2019; Ladd and Arvaniti, 2023), and is also a question of practical importance for speech synthesis (de Seyssel et al., 2024; Bauer et al., 2025). However no previous work has noticed this connection. While syntagmatically plausible — emphasis may be rendered clearer by contrast with previous reduction — and plausible also because of the existence of various pre-prominence prosodic tendencies (Roessig, 2024; Bishop and Kim, 2018; Breen et al., 2010), it is noteworthy that here, with no effort, the relationship appeared from the data.

Third, we followed up by examining this connection in 6 TED talks. We plotted the average level of reduction in the vicinity of emphasized words, with both emphasized stressed syllables and regular stressed syllables annotated. These were annotated using strict criteria, giving 68 instances of the former and 56 of the latter. We found that reduction was on average greater in the emphasized case over a region of about 400 to 200 milliseconds before the start of the word with stress.

Fourth, we examined Youtube videos by 10 vloggers, annotated for three degrees of prominence following the DIMA guidelines (Kügler et al., 2019). There were 741 instances of emphatic (level-3) reduction, 2680 of strong (level-2) reduction, and 1638 of weak (level-1). Figure 2 shows the average level of reduction over the 500 ms before the prominence peak. There is an apparent tendency to reduction over 500 to 280 ms before the peak. Incidentally, the steep dive in reduction before emphasized peaks is likely because the emphasized syllables themselves are often enunciated and lengthened.

As an additional follow-up, we examined each individual's pattern of pre-emphasis reduction. Six of the speakers had similar patterns, of which Speaker AD was an exemplar, as seen in Figure 3. To roughly gauge the strength of the connection for this speaker, we compared the average reduction over the 400–300ms before the peak for the emphatic prominences versus the weak ones, of which there were 53 and 166 occurrances, respectively. For this, the p value by a t-test is $< .0001$ (although we are not testing a hypothesis) and the effect size is about 0.71 standard deviations. However, the individual variation was great, with some tending to produce the reduction more weakly, more variably, earlier, or later, and with one speaker apparently lacking any tendency to pre-emphasis reduction.

Clearly we have raised more questions than we can answer, including how these patterns are aligned with segmental landmarks or other anchors, and we hope that future work will explore these. However, these illustrations do demonstrate how this new tool can support new methods of inquiry and enable new discoveries.

9. Summary

Our major contribution is the finding that a simple architecture — just a linear regression layer over HuBert features, trained on just a modest amount

of data — can predict perceptions of reduction. Its performance is in the range of, and sometimes outperforms, human inter-annotator agreement, and far outperforms previous models (Ward and Ortega, 2024).

Our second major contribution is the reduction estimation tools, for English and Spanish. These are a set of python scripts that invoke the HuBert model, train and/or apply a linear regression model, and so on. They also include a ready-to-use English model, which we call ReduEst, trained on the primary data, whose predictions correlate at 0.55 with average human judgments. This is released at `github/nigelgward/ISG_Reduction_Model`. This model can support statistical studies of reduction in corpora, and probably also other use cases, including better speech synthesizer training and evaluation.

Other contributions include the release of annotations to support testing other systems for this task, statistics on inter-annotator agreement, confirmation that reduction patterns are language-dependent, the identification of language properties that can complicate reduction estimation, and the discovery of a tendency to pre-emphasis reduction. In addition, we flag some priorities for future work, especially creating larger data resources and investigating how human perceptions of reduction differ.

This work opens the door to broader and deeper investigations of the functions of phonetic reduction in dialog. This should, in turn, enable more accurate estimation of user states and intents from speech, and more expressive speech synthesis. Ultimately the benefit will be robots, AI agents, and dialog systems able to communicate more effectively with users.

10. Acknowledgments

We thank Elina Banzina for the TED Talk annotations, Stephanie Berger for the Youtube DIMA annotations, and Marcel de Korte for comments. This work was supported in part by the National Science Foundation (Award 2348085), by the AI Research Institutes program of the NSF and the Institute of Education Sciences, U.S. Department of Education, through Award #2229873 (National AI Institute for Exceptional Education), by the Air Force Office of Scientific Research under award number FA9550-24-1-0281, and by the Otto Mønsted Fund.

11. Bibliographical References

Bistra Andreeva, William J Barry, and Jacques C Koreman. 2014. A cross-language corpus for studying the phonetics and phonology of prominence. In *LREC*, pages 326–330.

Tomás Arias-Vergara, Elizabeth Londoño-Mora, Paula A Pérez-Toro, Maria Schuster, Elmar Nöth, Juan Rafael Orozco-Arroyave, and Andreas Maier. 2023. Measuring phonological precision in children with cleft lip and palate. In *Interspeech*, pages 4638–4642.

Judith Bauer, Frank Zalkow, Meinard Müller, and Christian Dittmar. 2025. Explicit emphasis control in text-to-speech synthesis. In *Speech Synthesis Workshop (SSW)*, pages 21–27.

Grégory Beller, Nicolas Obin, and Xavier Rodet. 2008. Articulation degree as a prosodic dimension of expressive speech. In *Speech Prosody*, pages 681–684.

Jason Bishop and Boram Kim. 2018. Anticipatory shortening: Articulation rate, phrase length, and lookahead in speech production. In *9th International Conference on Speech Prosody*.

Mara Breen, Evelina Fedorenko, Michael Wagner, and Edward Gibson. 2010. Acoustic correlates of information structure. *Language and cognitive processes*, 25(7-9):1044–1098.

Rachel Steindel Burdin and Cynthia G Clopper. 2015. Phonetic reduction, vowel duration, and prosodic structure. In *International Congress of the Phonetic Sciences*.

Kaleigh Susan Cammenga. 2018. *The Effect of Vocal Fry on Speech Intelligibility*. Ph.D. thesis, Michigan State University.

Francesco Cangemi and Oliver Niebuhr. 2018. Rethinking reduction and canonical forms. In Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler, and Margaret Zellers, editors, *Rethinking Reduction*, pages 277–302. de Gruyter.

Lei Chen, Chenglin Jiang, Yiwei Gu, Yang Liu, and Jiahong Yuan. 2022. Automatically detecting reduced-formed English pronunciations by using deep learning. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 22–26.

Bronya R Chernyak, Talia Ben Simon, Yael Segal, Jeremy Steffman, Eleanor Chodroff, Jennifer S Cole, and Joseph Keshet. 2022. DeepFry: Identifying vocal fry using deep neural networks. In *Interspeech*, pages 3578–3582.

Jennifer Cole, José I Hualde, Caroline L Smith, Christopher Eager, Timothy Mahrt, and Ricardo Napoleão de Souza. 2019. Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, 75:113–147.

Maureen de Seyssel, Antony D'Avirro, Adina Williams, and Emmanuel Dupoux. 2024. EmphAssess : a prosodic benchmark on assessing emphasis transfer in speech-to-speech models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 495–507.

Patrick Cormac English, John D Kelleher, and Julie Carson-Berndsen. 2023. Discovering phonetic feature event patterns in transformer embeddings. In *Interspeech*.

Mirjam Ernestus and Natasha Warner. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics*, 39(SI):253–260.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520.

Joakim Gustafson, Eva Székely, Simon Alexandersson, and Jonas Beskow. 2023a. Casual chatter or speaking up? Adjusting articulatory effort in generation of speech and animation for conversational characters. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition*.

Joakim Gustafson, Éva Székely, and Jonas Beskow. 2023b. Generation of speech and facial animation with controllable articulatory effort for amusing conversational characters. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–9.

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.

Wen-Chin Huang, Benjamin Peloquin, Justine Kao, Changhan Wang, Hongyu Gong, Elizabeth Salesky, Yossi Adi, Ann Lee, and Peng-Jen Chen. 2023. A Holistic Cascade System, Benchmark, and Human Evaluation Protocol for Expressive Speech-to-Speech Translation. In *ICASSP*.

T Florian Jaeger and Esteban Buz. 2017. Signal reduction and linguistic encoding. *The Handbook of Psycholinguistics*, pages 38–81.

Yishan Jiao, Visar Berisha, Julie Liss, Sih-Chiao Hsu, Erika Levy, and Megan McAuliffe. 2016. Articulation entropy: An unsupervised measure of articulatory precision. *IEEE Signal Processing Letters*, 24(4):485–489.

Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William D Raymond. 1998. Reduction of English function words in Switchboard. In *International Conference on Spoken Language Processing*.

Greg Kochanski, Esther Grabe, John Coleman, and Burton Rosner. 2005. Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2):1038–1054.

Frank Kügler, Stefan Baumann, Bistra Andreeva, Bettina Braun, Martine Grice, Jana Neitsch, Oliver Niebuhr, Jörg Peters, Christine T. Röhr, Antje Schweitzer, and Petra Wagner. 2019. Annotation of German intonation: DIMA compared with other annotation systems. In *ICPhS*.

D Robert Ladd and Amalia Arvaniti. 2023. Prosodic prominence across languages. In *Annual Review of Linguistics*, volume 9, pages 171–193. Annual Reviews.

Harm Lameris, Joakim Gustafsson, and Éva Székely. 2023. Beyond style: Synthesizing speech with pragmatic functions. In *Interspeech*, pages 3382–3386.

Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, and Shrikanth Narayanan. 2006. A study of emotional speech articulation using a fast magnetic resonance imaging technique. In *Ninth International Conference on Spoken Language Processing*.

Heike Lehnert-LeHouillier, Susana Terrazas, and Steven Sandoval. 2020. Prosodic entrainment in conversations of verbal children and teens on the autism spectrum. *Frontiers in Psychology*, 11:2718.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2022. On the utility of self-supervised models for prosody-related tasks. In *IEEE Workshop on Spoken Language Technology (SLT)*, pages 1104–1111.

John Local and Gareth Walker. 2012. How phonetic features project more talk. *Journal of the International Phonetic Association*, 42:255–280.

Nichola Lubold, Stephanie A Borrie, Tyson S Barrett, Megan M Willi, and Visar Berisha. 2019. Do conversational partners entrain on articulatory precision? In *Interspeech*, pages 1931–1935.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, et al. 2022. Spoken language interaction with robots: Research issues and recommendations. *Computer Speech and Language*, 71. 101225.

Devang S Ram Mohan, Vivian Hu, Tian Huey Teh, Alexandra Torresquintero, Christopher G R Wallis, Marlene Staib, Lorenzo Foglianti, Jiameng Gao, and Simon King. 2021. Ctrl-p: Temporal control of prosodic variation for speech synthesis. In *Interspeech*, pages 3875–3879.

Jaya Narain, Vasudha Kowtha, Colin Lea, Lauren Tooley, Dianna Yee, Vikramjit Mitra, Zifang Huang, Miquel Espi Marques, Jon Huang, Carlos Avendano, et al. 2025. Voice quality dimensions as interpretable primitives for speaking style for atypical speech and affect. In *Interspeech*.

Oliver Niebuhr. 2014. 'A little more ironic': Voice quality and segmental reduction differences between sarcastic and neutral utterances. In *7th Speech Prosody Conference*, pages 608–612.

Oliver Niebuhr. 2016. Rich reduction: Sound-segment residuals and the encoding of communicative functions along the hypo-hyper scale. In *7th Tutorial & Research Workshop on Experimental Linguistics, ExLing*, pages 11–24. International Speech Communication Association (ISCA).

Oliver Niebuhr. 2017. Clear speech — mere speech? How segmental and prosodic speech reduction shape the impression that speakers create on listeners. In *Interspeech*, volume 18, pages 894–898.

Oliver Niebuhr, Karin Gors, and Evelin Graupe. 2013. Speech reduction, intensity, and F0 shape are cues to turn-taking. In *SigDial*, pages 261–269.

Benjamin Picart, Thomas Drugman, and Thierry Dutoit. 2014. Analysis and HMM-based synthesis of hypo and hyperarticulated speech. *Computer Speech & Language*, 28(2):687–707.

Leendert Plug. 2005. From words to actions: The phonetics of eigenlijk in two communicative contexts. *Phonetica*, 62(2-4):131–145.

Leendert Plug. 2011. Phonetic reduction and informational redundancy in self-initiated self-repair in Dutch. *Journal of Phonetics*, 39(3):289–297.

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, and Julien Pinquier. 2023. Can we use speaker embeddings on spontaneous speech obtained from medical conversations to predict intelligibility? In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Simon Roessig. 2024. The inverse relation of pre-nuclear and nuclear prominences in german. *Laboratory Phonology*, 15(1):1–43.

Ming Tu, Anna Grabek, Julie Liss, and Visar Berisha. 2018. Investigating the role of L1 in automatic pronunciation evaluation of L2 speech. In *Interspeech*, pages 1636–1640.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, et al. 2019. Speech synthesis evaluation: State-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop*.

Nigel G Ward. 2019. *Prosodic Patterns in English Conversation*. Cambridge University Press.

Nigel G Ward, Raul O Gomez, Carlos A Ortega, and Georgina Bugarini. 2025a. Phonetic reduction is associated with positive assessment and other pragmatic functions. *Speech Communication*, 175. 103305.

Nigel G Ward, Divette Marco, and Olac Fuentes. 2025b. Which prosodic features matter most for pragmatics? *IEEE ICASSP*.

Nigel G Ward and Carlos A Ortega. 2024. Preliminaries to a study of the pragmatic functions of reduced articulatory precision in dialog. Technical Report UTEP-CS-24-23, University of Texas at El Paso, Department of Computer Science.

Nigel G. Ward, Steven D. Werner, David G. Novick, Tatsuya Kawahara, Elizabeth E. Shriberg, Louis-Philipp Morency, and Catharine Oertel. 2013. The similar segments in social speech task. In *MediaEval Workshop*.

Duane G Watson. 2010. The many roads to prominence: Understanding emphasis in conversation. In B. Ross, editor, *Psychology of Learning and Motivation*, pages 163–183. Elsevier.

Shu-Wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, et al. 2024. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899.

Margaret Zellers, Barbara Schuppler, and Meghan Clayards. 2018. Introduction, or: Why rethink reduction? In Francesco Cangemi, Meghan Clayards, Oliver Niebuhr, Barbara Schuppler, and Margaret Zellers, editors, *Rethinking Reduction*, pages 1–24. de Gruyter.

12. Language Resource References

John J. Godfrey and Edward C. Holliman and Jane McDaniel. 1992. *Switchboard: Telephone speech corpus for research and development*. Linguistic Data Consortium, Release 2, ISLRN 988-076-156-109-5.