ADEPT

ADEPT: A Dataset for Evaluating Prosody Transfer

Alexandra Torresquintero[1], *Tian Huey Teh*[1], *Christopher G. R. Wallis*[1], *Marlene Staib*[1], *Devang S Ram Mohan*[1], *Vivian Hu*[1], *Lorenzo Foglianti*[1], *Jiameng Gao*[1], *Simon King*[1,2]

[1]Papercup Technologies Ltd., United Kingdom
[2]University of Edinburgh, United Kingdom
{alexandra,tian,chris}@papercup.com

Abstract

Text-to-speech is now able to achieve near-human naturalness and research focus has shifted to increasing expressivity. One popular method is to transfer the prosody from a reference speech sample. There have been considerable advances in using prosody transfer to generate more expressive speech, but the field lacks a clear definition of what successful prosody transfer means and a method for measuring it. We introduce a dataset of prosodically-varied reference natural speech samples for evaluating prosody transfer. The samples include global variations reflecting emotion and interpersonal attitude, and local variations reflecting topical emphasis, propositional attitude, syntactic phrasing and marked tonicity. The corpus only includes prosodic variations that listeners are able to distinguish with reasonable accuracy, and we report these figures as a benchmark against which text-to-speech prosody transfer can be compared. We conclude the paper with a demonstration of our proposed evaluation methodology, using the corpus to evaluate two text-to-speech models that perform prosody transfer.
Index Terms: TTS prosody transfer, evaluation

1. Introduction

Text-to-speech (TTS) research relies on subjective human evaluations. There are well-established methods to assess naturalness – A/B comparisons, Mean Opinion Score (MOS), or MUSHRA [1] – and intelligibility using a transcription task. But, now that TTS routinely matches human speech intelligibility and approaches human naturalness, focus has shifted to expressivity. Perhaps the most popular current method is to transfer prosody from an expressive reference sample.

Subjective methods for evaluating prosody transfer are not well developed. Some provide listeners with a reference and measure how well its prosody was transferred. [2, 3, 4] employed an AXY discrimination task in which listeners judge how similar generated samples X and Y are to reference A; [5] used this task with trained linguists. [6] claimed that successful transfer of a song melody to speech indicates successful prosody transfer. Other methods provide no reference. [7] use a preference test. [3, 4, 8] assume that measuring naturalness with MOS is sufficient. There is no common method, which hinders the comparison of prosody transfer approaches. Therefore, we release a corpus (DOI: 10.5281/zenodo.5117102) of expressive natural speech reference samples that can be used within our proposed evaluation methodology.

The goal of prosody transfer is to synthesise a given text with the prosody of a reference utterance. By varying the reference, the system generates prosodically-distinct renditions. Our proposed evaluation method therefore starts from a corpus of English sentences, each with multiple prosodically-distinct natural renditions. The TTS system under evaluation is required to transfer the prosody from a reference utterance taken from this corpus, and the evaluation metric is the accuracy with which listeners perceive the correct prosody.

To identify the effectiveness of the transfer across different aspects of prosody, the natural utterances fall into prosodic classes within which listeners are able to perform a categorisation task. To find suitable classes, we examine several spoken phenomena known to have prosodic consequences (§2). Within each of these, we identify subcategories reported in the literature to have perceptually distinct prosody (§3.1). The listeners' task will then be to categorise a synthetic rendition as the correct subcategory. After recording the natural utterances (§3.2), we confirm that listeners can perform the categorisation task on natural speech (§3.3), and discard subcategories and utterances that listeners cannot reliably categorise. We finish with a demonstration of our proposed method using synthetic speech (§4), and a discussion and suggestions for further work (§5).

2. Speech classes with prosodic effect

Prosody has many definitions, but we adopt [9, p. 196]: high-level structures that account for F_0, duration, amplitude, spectral tilt, and segmental reduction patterns in speech. These structures have local and global effects [10]. Many aspects of speech are part of prosody by this definition.

[10] describes two phenomena, emotion and attitude, that speakers express through F_0, amplitude, duration, and spectral tilt prosodic cues [10, 11]. **Emotion** is an inner state of the speaker (e.g., joy), whilst attitude is towards something external. **Interpersonal attitude** is toward the listener, e.g., friendliness. **Propositional attitude** is toward what is being said, e.g., incredulity. Emotion and interpersonal attitude have a global prosodic effect, and propositional attitude has a local effect [10].

Topical emphasis occurs when the topic is prosodically highlighted because of its relative importance to other words in the sentence, such as *not* in *I will NOT go*. It has local effects on F_0, amplitude, and duration [12, p. 15].

Syntactic phrasing affects prosody through perceivable intonation groups: the end of a phrase exhibits lengthening of the phrase-final word and following pause within a sentence [13].

In English speech there will always be a syllable that carries the greatest lexical stress across the sentence [14][15]; we call this phenomenon **marked tonicity**. It has similar, though subtler, prosodic effect on prominent words as topical emphasis, but can also cause segmental reduction on non-stressed words.

We have introduced 2 classes that have a global prosodic effect (emotion and interpersonal attitude), and 4 classes that have local effect (propositional attitude, topical emphasis, syntactic phrasing, and marked tonicity). Other structures have similar prosodic effects, such as style (whispered, instructional, broadcasting, etc.) and speaker identity (age, gender, accent, etc.). However, these are less likely to change per-utterance, so are less relevant to most TTS prosody transfer use cases.

3. Dataset and evaluation design

These 6 classes are not mutually exclusive; in one sentence a speaker can sound sad (emotion) and polite (interpersonal attitude) and emphasise a word. But to use these classes to evaluate TTS prosody transfer, we require that their prosodic effects are perceivable in isolation. We propose a disambiguation task in which listeners are asked to categorise speech samples based only on their prosody. For example, the sentence *It's snowing* can be said both happily or sadly. If listeners who are played both recordings can correctly identify which is sad and which is happy, we can conclude listeners can perceive emotion based on prosody alone. In the following sections, we describe how we used prior research to determine suitable ambiguity for such a disambiguation task (§3.1), the design and recording of the natural speech from which prosody will be transferred (§3.2), pretests to find the most reliable task design and data for evaluating prosody transfer (§3.3), and our final proposed evaluation methodology (§3.4).

3.1. Perceivable subcategories or interpretations

For each of the 6 classes, listeners will perform a disambiguation task using prosody: therefore, we needed to identify prosodic ambiguity within each class. For emotion, interpersonal attitude, propositional attitude, and topical emphasis classes, we found suitable ambiguity in *subcategories* of the class. For syntactic phrasing and marked tonicity, sentences with two *interpretations* had suitable ambiguity.

For **emotion**, [16] report the perceivability of anger, disgust, fear, sadness, happiness, pleasant surprise, and neutral, in 4 languages. We discarded pleasant surprise as it was 1 in 3 of their perceptually-invalid items. We renamed happiness to joy which is less likely to be confused for something more complex like nostalgia. This left 5 perceptually distinct subcategories of the emotion class: anger, disgust, fear, sadness, and joy.

[17, 18] measured the perceivability (in Brazilian Portuguese) of 12 **interpersonal attitudes**. Listeners had to disambiguate arrogance, authority, contempt, irritation, politeness, seduction, and neutral in question and statement utterances. We eliminated subcategories whose perceivability interacted with speaker gender (arrogance, irritation, seduction), and subcategories confused with neutral (contempt statements, polite questions). We eliminated authoritative questions because authoritative statements were perceived more strongly. This left three subcategories of the interpersonal attitude class: contemptuous questions, authoritative statements, and polite statements.

[17, 18] also measured **propositional attitude** by asking listeners to disambiguate between four question attitudes (rhetoricity, confirmation, incredulity, surprise) plus neutral, and five statement attitudes (irony, incredulity, surprise, doubt, obviousness) plus neutral. We eliminated rhetoricity questions for being confused with neutral, surprise questions for being confused with incredulity, and incredulity statements for being confused with irony. We renamed irony to sarcasm for clarity. This left 6 subcategories of the propositional attitude class: obviousness, surprise, sarcasm, and doubt statements, and incredulity and confirmation questions.

[19] show that acoustic differences can arise depending on the locus of **topical emphasis** in the sentence, yielding three subcategories: beginning, middle, and end.

[20] show that listeners can use prosodic cues to disambiguate meanings of a sentence with phrasing ambiguity. For example, *Put the dog food in the bowl on the floor* has two interpretations: put the dog food into the bowl that is on the floor, or put the dog food that is in the bowl onto the floor. The former meaning can be conveyed with a pause after *food*, and the latter with a pause after *bowl*. We used sentences with this **syntactic phrasing** ambiguity for our disambiguation task. These are not subcategories per se, so we refer to them as interpretations.

Sentences with part of speech ambiguity can be disambiguated by **marked tonicity** prosodic cues. [21, p. 55] gives an example sentence *He ate a little pudding* which has two interpretations: 1) he didn't eat very much pudding, in which 'a little' is a determiner to 'pudding', and the strongest lexical stress falls on 'pudding'; 2) he ate a small pudding, where it falls on 'little', which is an adjective for 'pudding'.

For each of the 6 classes, the next step was to design sentences that can be read in prosodically distinct ways by subcategory (of emotion, interpersonal attitude, propositional attitude, topic emphasis) or interpretation (for syntactic phrasing, marked tonicity), and record them with voice actors.

3.2. Sentence design and recording

For each class, we devised at least 20 sentences that could be spoken to express all the subcategories of, or two interpretations per sentence within, that class. For example, *"Look at that puppy."*, can be said in all five subcategories of the emotion class. Many of the sentences for syntactic phrasing and marked tonicity came from [22].

In addition to the sentence to be spoken, we devised a contextual cue to elicit the prosodically-distinct rendition portraying each subcategory/interpretation. For **emotion**, **interpersonal attitude**, and **propositional attitude**, the contextual cues were situations that evoke the subcategory, such as a teacher-student relationship implying an authoritative interpersonal attitude. For **topical emphasis**, the emphasised word was capitalised in the script, and contextual cues were wh-questions that implied the emphasis. For example, for the sentence *"Dogs play FETCH in parks"* with target emphasis on FETCH, the contextual wh-question was "Dogs play WHAT in parks?" For **syntactic phrasing** and **marked tonicity**, the contextual cues were paraphrases that made the intended meaning clear, such as the two explicit interpretations of the sentence *Put the dog food in the bowl on the floor* in §3.1. When possible, each sentence was also recorded in a 'neutral' style: no subcategory was expressed. As one of the two interpretations is required for each syntactic phrasing and marked tonicity sentence, a neutral style for these classes does not exist. Two voice actors, male and female, read the 552 sentences in Standard Southern British English.

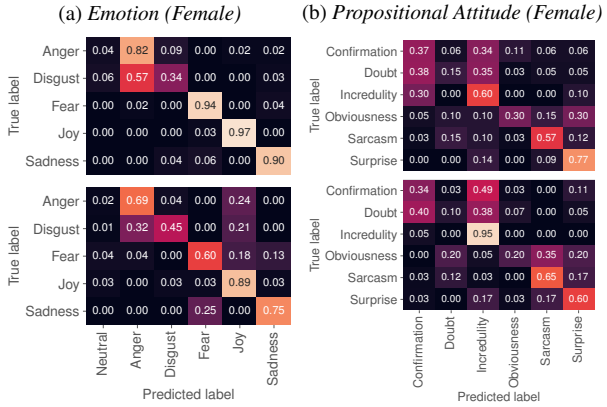
3.3. Pretests

Before using these sentences and subcategories/interpretations to evaluate prosody transfer, we tested that listeners consistently found them prosodically distinguishable in natural speech, and then selected the most appropriate task for them to perform.

We trialled various disambiguation task designs for each class (§3.3.2), then selected the most reliable design, and finally used that task to select the most consistently disambiguated sentences and subcategories for each speaker/class pair (§3.3.3). The resulting task design and selected sentences will be used in the final evaluation methodology (§3.4).

3.3.1. Listeners

Amazon Mechanical Turk was used to recruit 10 self-reported native English speakers who passed a short transcription test, per pretest. To filter listeners who could hear prosodic differences, we disqualified participants who: 1) across all questions,



(a) *Emotion (Female)* (b) *Propositional Attitude (Female)*

(c) *Emotion (Male)* (d) *Propositional Attitude (Male)*

Figure 1: *Confusion matrices for the top 5 sentences from classes with subcategories eliminated in pretests.*

selected an option significantly above chance (e.g., always selecting A); or 2) for a given correct subcategory, did not select any subcategory significantly above chance. E.g., for all questions whose correct answer is disgust, people who statistically significantly selected angry (or any other subcategory) were not disqualified. As 10 is a small sample size, we treated answers as binomial distributions with statistical significance at the 99% confidence interval.

3.3.2. Task design

As with the design of any listening evaluation, there were more potential design options than could be fully explored. Because we would subsequently be eliminating weak stimuli (§3.3.3) which would further improve recognition accuracy, in this task design phase we used our best judgement to make design decisions, but acknowledge that this will not always be optimal.

We assume task design is speaker-independent, so trialled multiple task designs only on the female speaker. The neutral stimulus, when it existed, was included as a sample unless otherwise specified. We considered two design choices: 1) Should we ask *directly* or *indirectly* about the subcategory/interpretation? E.g., do we directly ask which word carries most lexical stress, or present the sample in a context where that stress pattern would be preferred? 2) Should listeners categorise a *single stimulus* or choose which one of *multiple stimuli* matches a label?

For **emotion**, we tried a direct question. The multiple stimulus design asked listeners 'In which of the following samples does the speaker sound most *x*?' where *x* was one of disgusted, angry, joyful, fearful, or sad. The stimuli were samples of the same sentence in each of the subcategories plus neutral. The single stimulus design asked listeners 'Which emotion is most reflected in the speakers voice?', with six choices: disgust, anger, joy, fear, sadness, or none of these. Across all 20 sentences, we found higher recognition accuracy for each subcategory in the multiple stimulus task.

For **interpersonal attitude**, we tried a direct question multiple stimulus design. We asked 'Which of the following samples sounds most like *x*?' where *x* was an authoritative statement, a contemptuous question, or a polite statement.

For **topical emphasis**, we tried a single stimulus design and compared direct and indirect questions. The direct question asked listeners 'Which word is most strongly emphasised in the sample?', with three choices of the content words in the beginning, middle, and end of the sentence. The indirect question asked 'Which question is best answered by the sample?' The three choices were the context cues described in §3.2. Neutral was not included because this is not a correct response to any of the context cues. After disqualifying participants, recognition accuracies across all sentences for each subcategory were higher for the indirect design.

We tried an indirect question multiple stimulus design for **propositional attitude**, but excluded neutral because we believed 7 samples (6 subcategories + neutral) was too many to compare at once. Participants were asked 'Which audio fits best into the context: *x*?' where *x* was one of:

"Obviously _____."
"(Surprised) Wow! _____!"
"(Sarcastically) Well _____."
"(Unsure) Perhaps _____."
"Really? _____?"
"I just want to confirm that _____?"

For **syntactic phrasing**, we tried single stimulus and multiple stimulus designs with an indirect question. Listeners were asked 'Which is a better paraphrase of the sample?' in the single stimulus design, and 'Which sample fits the paraphrase best? *x*' in the multiple stimulus design, where *x* was a paraphrase from §3.2. The single stimulus design provided better results, potentially because it is easier to see the two alternative interpretations by reading two paraphrases. We also used this paraphrase single stimulus design for **marked tonicity**.

3.3.3. Elimination of weak stimuli

The best designs above for each class were re-run using male speaker stimuli. We then used the qualified participants to identify the five sentences for each speaker and class that had the highest recognition accuracy across all subcategories/interpretations. We eliminated any subcategories with less than 60% recognition accuracy in these sentences, assuming anything below this threshold was not perceivable enough to measure TTS against in the final evaluation method (§3.4).

Disgust was discarded as a subcategory of **emotion** because its recognition accuracy was less than 60% for both speakers (Figures 1a and c). All **interpersonal attitude** (Table 1) and **topical emphasis** subcategories met the 60% threshold. For

Table 1: *Interpersonal attitude pretest results*

	authority	contempt	politeness
female	60%	83%	85%
male	93%	60%	71%

topical emphasis, recognition accuracy was 100% for all subcategories and speakers. Figures 1b and d show that confirmation, doubt, and obviousness **propositional attitudes** did not meet the 60% threshold for either speaker, nor did sarcasm for the female speaker. For the 2 classes with sentence-dependent interpretations, we report accuracy per speaker for all top 5 sentence stimuli together, because subcategories of these sentences do not exist. For **syntactic phrasing** this accuracy was 90% for both speakers. For **marked tonicity**, this was 79% and 83% for the female and male stimuli respectively.

3.4. The proposed evaluation methodology

The final ADEPT evaluation methodology consists of 12 disambiguation tasks: 6 classes × 2 speakers. Each task uses 5 sentences with multiple prosodic renditions. Each task has one question per sentence and distinguishable subcategory or interpretation. As shown in Table 2, for each question there is one choice per subcategory or interpretation, plus neutral if it

Table 2: *Number of choices per question for each disambiguation task, whether questions are single or multiple stimulus, and whether neutral samples are included as stimuli.*

class	choices F	choices M	audio stimuli	neutral included
emotion	5	5	multiple	yes
interpersonal attitude	4	4	multiple	yes
topical emphasis	3	3	single	no
propositional attitude	3	4	multiple	yes
synctactic phrasing	2	2	single	-
marked tonicity	2	2	single	-

is present. For example, the final female propositional attitude test is a 10 question multiple stimulus task, and each question's three choices are the incredulity, surprise, and neutral samples.

For the final setup for propositional attitude, we include neutral because some subcategories were eliminated, and the incredulity context is updated to "(Incredulous) Really? _____?"

4. Evaluating TTS prosody transfer models

We demonstrate the use of the ADEPT evaluation methodology to compare synthetic speech generated by two recently-proposed TTS models that perform prosody transfer. At the same time, we establish a benchmark based on the recognition accuracy of natural speech.

Our two models are both based on a multi-speaker variant of Tacotron 2 [23, 24]. Following [2] we extend the Tacotron 2 architecture by adding a reference encoder that learns a fixed-length prosody embedding from the reference in unsupervised fashion (henceforth **Tacotron-Ref**). We compare this to **Ctrl-P** [25], which explicitly models three acoustic correlates of prosody (F_0, energy, and duration) per-phone. For supervised training, ground-truth values are extracted from the force-aligned training data. Each feature is normalised to zero mean and unit standard deviation, per speaker. During inference, values are extracted from the reference speech and normalised using the target speaker's train set statistics. This variable-length prosody representation is concatenated with the Tacotron 2 encoder output and attended over by the decoder.

Our training data comprised 24 h of non-fiction audiobook readings by the female speaker from the LJSpeech corpus [26], 20 h of fiction audiobook readings by the female speaker from the 2013 Blizzard Challenge [27], and 3 h of proprietary data (not from ADEPT) in order to include a male speaker.

We trained one **Ctrl-P** model and one **Tacotron-Ref** model on this corpus. Female samples were generated with LJ as the target speaker with prosody transferred from samples of the female ADEPT speaker. Male samples were generated with our proprietary speaker as the target and prosody transferred from samples of the male ADEPT speaker. The Griffin-Lim [28] algorithm was assumed to be sufficient for the current demonstration of the ADEPT evaluation methodology, but neural-vocoded samples could also have been used.

The evaluations on the natural speech and generated samples were performed on Amazon Mechanical Turk by self-reported native English speakers who passed a short English transcription test. In the pretests (§3.3.1) using only natural speech, we disqualified participants who couldn't hear any prosody. Since TTS cannot guarantee successful prosody transfer, this is no longer appropriate. Instead, we employed 5 'trap' questions. For multiple stimulus designs, trap questions required the listener to identify the sample that sounded most like 'English speech', where all audio files but one were time-reversed. For the single stimulus designs, trap questions replaced all options except the correct one with a description obviously unrelated to the sample. Participants who got any trap question wrong were disqualified and excluded from results. Each test had exactly 30 qualifying participants, as recommended by [29]. Results are shown in Table 3, with the 2 global classes above and the 4 local classes below.

Table 3: *Recognition accuracy (%) for each of a class' subcategories, or entire class that doesn't have subcategories, for female (F) and male (M) natural speech (N), and LJSpeech (LJ) and proprietary (Prop.) synthetic speakers from Ctrl-P (C) and Tacotron-Ref (T). Accuracies statistically significantly above chance are in bold (one-tailed binomial test; $p \leq 0.05$).*

		F	LJ		M	Prop.	
class	subcategory	N	C	T	N	C	T
emotion	anger	**95**	**31**	17	**83**	6	6
	fear	**80**	16	**40**	**52**	20	9
	joy	**90**	**33**	18	**88**	**75**	**54**
	sadness	**88**	**49**	21	**62**	**53**	13
inter-personal attitude	authority	**47**	14	26	**60**	**35**	29
	contempt	**49**	**50**	29	**52**	**35**	30
	politeness	**37**	23	26	**63**	**37**	29
topical emphasis	beginning	**87**	**68**	**52**	**82**	**87**	**66**
	middle	**79**	**67**	**43**	**69**	**71**	33
	end	**70**	**67**	**45**	**62**	**63**	32
propo-sitional attitude	incredulity	**63**	40	40	**71**	**75**	**33**
	sarcasm	-	-	-	**62**	**48**	**49**
	surprise	**73**	32	33	**66**	**72**	**47**
syntactic phrasing		**84**	**77**	**66**	**80**	**84**	**81**
marked tonicity		**74**	**69**	**58**	**62**	**58**	48

The ADEPT evaluation measures the success of prosody transfer for each model-voice combination on each subcategory, or class for classes without subcategories. Both Ctrl-P and Tacotron-Ref perform better than chance in some cases, with Ctrl-P doing so more and matching the natural speech benchmark for several classes.

5. Discussion and Conclusion

As expected, accuracies for natural speech are all significantly above chance, albeit lower than in pretests, probably as a result of different qualifying criteria. Beyond model comparison, the ADEPT evaluation methodology also enables a host of other analyses. For instance, one could investigate differences in model performance between local and global prosodic classes, examine within-class confusion matrices, or compare performance transferring prosody from different source voices.

In this work, we introduced six high level local and global prosodic classes of speech that can be used in disambiguation tasks to evaluate TTS prosody transfer. This evaluation methodology allows researchers to both compare performance of their models against each other, and against a natural benchmark of target performance. Further work might consider if these classes and their subcategories/interpretations are viable for a cross-lingual prosody transfer application, or if they can be used to evaluate prosody in TTS in general.

6. Acknowledgements

We thank our adviser Mark Gales for guidance and our voice actors Nishad and Laura who read all sentences even when they didn't make sense.

7. References

[1] "Method for the subjective assessment of intermediate quality level of coding systems. ITU recommendation ITU-R BS.1534-1," *International Telecommunication Union Radiocommunication Assembly*, 2003.

[2] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," *PMLR*, vol. 80, pp. 4693–4702, Jul. 2018.

[3] E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," Oct. 2019. [Online]. Available: arXiv:1906.03402

[4] S. Gururani, K. Gupta, D. Shah, Z. Shakeri, and J. Pinto, "Prosody transfer in neural text to speech using global pitch and loudness features," May 2020. [Online]. Available: arXiv:1911.09645

[5] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *INTERSPEECH*, pp. 4387–4391, Oct. 2020.

[6] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," *ICASSP*, pp. 5911–5915, May 2019.

[7] T. Kenter, M. Sharma, and R. Clark, "Improving the Prosody of RNN-Based English Text-To-Speech Synthesis by Incorporating a BERT Model," *INTERSPEECH*, pp. 4412–4416, Oct. 2020.

[8] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," *ICASSP*, pp. 6264–6268, May 2020.

[9] S. Shattuck-Hufnagel and A. Turk, "A prosody tutorial for investigators of auditory sentence processing," *Journal of Psycholinguistic Research*, vol. 25, no. 2, pp. 193–247, Apr. 1996.

[10] J. A. de Moraes, "From a prosodic point of view: Remarks on attitudinal meaning," in *Pragmatics and Prosody: Illocution, Modality, Attitude, Information Patterning and Speech Annotation*. Firenze University Press, 2011, pp. 19–37.

[11] A. Rilliard, D. Erickson, T. Shochi, and J. A. de Moraes, "Social face to face communication – American English attitudinal prosody," *INTERSPEECH*, pp. 1648–1652, Aug. 2013.

[12] D. Bolinger, *Intonation and Its Parts: Melody in Spoken English*. Stanford University Press, 1985.

[13] D. W. Allbritton, G. McKoon, and R. Ratclif, "Reliability of prosodic cues for resolving syntactic ambiguity," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 3, pp. 714–745, 1996.

[14] M. Mateo, "Exploring pragmatics and phonetics for successful translation," *Vigo International Journal of Applied Linguistics*, no. 11, pp. 111–135, 2014.

[15] J. C. Wells, "Tonicity: Where does the nucleus go?" in *English Intonation: An Introduction*. Cambridge University Press, Aug. 2006, pp. 93–186.

[16] M. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. Kotz, "Factors in the recognition of vocally expressed emotions: A comparison of four languages," *Journal of Phonetics*, vol. 37, pp. 417–435, 2009.

[17] J. A. de Moraes, A. Rilliard, B. A. de Oliveira Mota, and T. Shochi, "Multimodal perception and production of attitudinal meaning in Brazilian Portuguese," *Speech Prosody*, May 2010.

[18] J. A. de Moraes, A. Rilliard, D. Erickson, and T. Shochi, "Perception of attitudinal meaning in interrogative sentences of Brazilian Portuguese," *ICPhS*, pp. 1430–1433, Aug. 2011.

[19] S. J. Eady and W. E. Cooper, "Speech intonation and focus location in matched statements and questions," *The Journal of the Acoustical Society of America*, vol. 80, no. 2, pp. 402–415, 1986.

[20] T. Kraljic and S. E. Brennan, "Prosodic disambiguation of syntactic structure: For the speaker or for the addressee?" *Cognitive Psychology*, vol. 50, no. 2, pp. 194–231, 2005.

[21] D. Hirst, "Stress," in *Intonative features: A syntactic approach to English intonation*. The Hague: Mouton, 1977, pp. 55–74.

[22] P. J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *The Journal of the Acoustical Society of America*, vol. 90, no. 6, pp. 2956–2970, 1991.

[23] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning Wavenet on mel spectrogram predictions," *ICASSP*, pp. 4779–4783, Apr. 2018.

[24] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. J. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *INTERSPEECH*, pp. 2080–2084, Sep. 2019.

[25] D. S. R. Mohan, V. Hu, T. H. Teh, A. Torresquintero, C. G. R. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King, "Ctrl-P: Temporal control of prosodic variation for speech synthesis," *INTERSPEECH*, 2021.

[26] K. Ito and L. Johnson, "The LJ speech dataset," 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[27] Lessac Technologies Inc., "Voice Factory audiobook recordings for Blizzard 2013," 2013.

[28] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[29] M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations," *INTERSPEECH*, pp. 3476–3480, Sep. 2015.