Implementing acoustic-prosodic entrainment in a conversational avatar

Rivka Levitan [1], Štefan Beňuš [2,3], Ramiro H. Gálvez [4], Agustín Gravano [4,5], Florencia Savoretti [4], Marian Trnka [3], Andreas Weise [1], Julia Hirschberg [6]

[1] Department of Computer and Information Science, Brooklyn College CUNY, USA
[2] Constantine the Philosopher University in Nitra, Slovakia
[3] Institute of Informatics, Slovak Academy of Sciences, Slovakia
[4] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina
[5] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina
[6] Department of Computer Science, Columbia University, USA

levitan@sci.brooklyn.cuny.edu, sbenus@ukf.sk, rgalvez@dc.uba.ar, gravano@dc.uba.ar, ssavoretti@dc.uba.ar, andreas.weise1204@gmx.de, Marian.Trnka@savba.sk, julia@cs.columbia.edu

Abstract

Entrainment, aka accommodation or alignment, is the phenomenon by which conversational partners become more similar to each other in behavior. While there has been much work on some behaviors there has been little on entrainment in speech and even less on how Spoken Dialogue Systems which entrain to their users' speech can be created. We present an architecture and algorithm for implementing acoustic-prosodic entrainment in SDS and show that speech produced under this algorithm conforms to the feature targets, satisfying the properties of entrainment behavior observed in human-human conversations. We present results of an extrinsic evaluation of this method, comparing whether subjects are more likely to ask advice from a conversational avatar that entrains vs. one that does not, in English, Spanish and Slovak SDS.
Index Terms: entrainment, alignment, virtual agents, avatars

1. Introduction

Entrainment, also termed *accommodation* or *alignment*, is the phenomenon by which conversational partners become more similar to each other in their behavior. Originally studied by cognitive psychologists, entrainment has been of great interest in recent years to computer scientists, who have used objective entrainment measures and corpus-based approaches to study the phenomenon in human-human and human-computer dialogues; these studies have shown that entrainment is associated with aspects of dialogue quality, including likability and task success. There has been some work on Spoken Dialogue Systems (SDS) that can dynamically entrain to a human user's language. However, there is almost no research on how SDS might entrain to the acoustic-prosodic features of a user's speech, and how this might affect dialogue quality.

There are two reasons for creating acoustic-prosodic entrainment in SDS. First, the general aim in SDS is to produce system behavior as close to human behavior as possible. Since humans entrain to their conversational partners in acoustic-prosodic features, this view indicates that SDS should as well. Second, entrainment might increase a user's perception of a system's quality in the same way that has been found to increase dialogue quality in human-human conversations [1, 2, 3, 4, 5].

This work presents an architecture for implementing acoustic-prosodic entrainment in an SDS. We show that an entrainment algorithm implemented under this architecture generates speech with the desired acoustic-prosodic properties. This approach is modular and self-contained and thus easily inserted into the structure of a pre-existing system [6]. We present results of an extrinsic evaluation, investigating whether subjects are more likely to ask advice from an entraining avatar than from one that does *not* entrain. We report three pilot studies in American English, Spanish and Slovak that provide evidence that entraining systems are a promising direction for SDS.

2. Related work

Empirical evidence of entrainment in human-human conversations has been documented for acoustic-prosodic features such as intensity [7, 8, 9], speaking rate [2], and pitch [8, 9]. Humans have been shown to entrain to their interlocutor's language at the lexical [10] or syntactic [11, 12] level, and on linguistic style [13, 14, 15]. Motivated by Communication Accommodation Theory (CAT) [16], which holds that speakers converge to or diverge from their interlocutors in order to attenuate or accentuate social differences, many studies have found links between entrainment and positive social behavior: entraining conversational partners are perceived as more socially attractive, more competent, more likable, interactions with them as more successful, entrainment is positively correlated with learning gains in automatic tutoring system and task success in Map Tasks [17, 18, 5, 3, 19, 1, 20, 21].

Despite strong interest in acoustic-prosodic entrainment, there has been no research on the scenario in which the computer entrains on acoustic-prosodic features, with notable exceptions for [6] on system entrainment on speaking rate and intensity and, subsequently, [22]'s adoption of this approach for entrainment on pitch in a tutoring system, in which pitch adaptation was shown to be associated with naturalness and rapport. There has been more research on *lexical* entrainment by a computer (e.g. [23, 24]), but since lexical features are discrete, and the technology involved is natural language generation, these approaches are qualitatively different from the proposed work, which addresses entrainment on continuous acoustic-prosodic features, and involves text-to-speech (TTS) technologies.

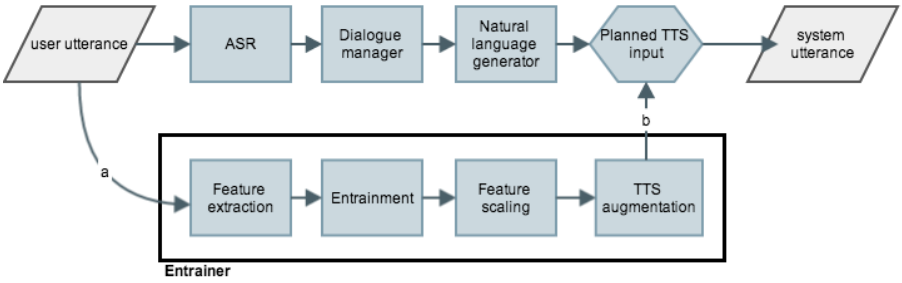

Figure 1: The entrainer integrated in an existing system.

3. Implementing entrainment

Entrainment relies on two processes: (1) Perceiving features of the interlocutor's speech, and (2) reproducing those features in the speech output. For (1), we use a Praat [25] script to extract acoustic-prosodic features from a user utterance. We can then accomplish (2) in two ways: by transforming those features into SSML markup for a TTS output utterance, or by transforming the acoustic-prosodic properties of the output once it has been synthesized. Our method can thus be easily integrated into an existing system, as shown in Figure 1, which depicts the SSML method, sending the user utterance to the feature extraction script as it goes through Automatic Speech Recognition (ASR) (arc `a`) and then augmenting the planned TTS input with markup tags output by the entrainer (arc `b`).

3.1. Feature extraction

Entrainment is triggered when the system captures a user utterance. While ASR translates the audio to text, a Praat script extracts acoustic-prosodic features from the signal, calculating mean intensity in dB and f0 in Hz. Other features, such as jitter, shimmer, or NHR could also easily be extracted. Speaking rate can be calculated from the speech signal (e.g. by Prosogram [26] or AuToBI [27]), or by counting syllables in the ASR output. Feature extraction takes less time than ASR and so does not add to system response latency.

3.2. Entrainment

Once prosodic features of the input are identified, the system must decide how to adapt its own output. Production is linked to perception [3], but in human-human conversations entrainment is affected by many factors, including differences in vocal tract length or utterance pragmatics, temporal factors, the relationship between speakers, and gender [2, 1, 28, 29] and operates differently for different prosodic features, languages and turn types [28, 30]. Here we simply match the target output to the input for each feature we wish to entrain on.

Before synthesizing TTS output with entrained parameters, each attribute value is compared against the range of acceptable values. Generating these ranges requires a configuration step, in which the range resulting in utterances perceived as normal (and not, for example, as unusually loud, slow, or distorted) are identified. If a value lies outside this range, it is replaced with the closest maximum value. So, we modify the goal of absolute entrainment to ensure natural-sounding output. In the case in which the user's input volume falls below the normal range, instead of producing the utterance inaudibly, the system outputs each utterance in the range closest to the extracted feature values. Some form of entrainment is still observable: the *entrained* mean is closer to the user's (perceived) mean than a *random* intensity level. Even if the extracted feature values are far from the user's, the system has performed no worse than a non-entraining system would: it has effectively fallen back on a standard behavior for this feature.

3.3. Adaptation

The system output can be adapted in one of two ways to match the extracted user feature values, either by appropriate SSML markup or by manipulation of the default output through Praat. We used MaryTTS [31], version 5.2, for testing and found that latency and accuracy depend both on the order in which individual features are adapted and on the method used.

Manipulation of the speech rate can have a notable effect on pitch with both methods but much more so with Praat than with SSML (10s of Hertz in some cases compared to less than 10). It can also take up to twice as long to use Praat. Pitch adaptation impacts intensity, more so with SSML (+-3dB) than with Praat (+-1dB). However, if done through SSML it causes no measurable change in latency compared to synthesis without markup for pitch, whereas it is "costly" with Praat. Lastly, intensity cannot be adapted for MaryTTS, as it currently does not support the `volume` keyword; this points to a more general problem with the SSML approach, that it depends on individual TTS implementations, while, with Praat, it is simple and fast.

In our experiments with an entraining avatar in English, Spanish, and Slovak (described in more detail in Section 5), we tested the accuracy of entrainment using SSML markup for two unit-selection text-to-speech engines, Cepstral (English) and MaryTTS (Spanish), and one parametric TTS implemented on the open-source engine HTS (Slovak).

For all three TTS, correlations between the target speech rates and the actual speech rates of the entrained TTS output were strong (over 0.70) and statistically significant. Entrainment on intensity was only tested for the Cepstral TTS; correlations were almost perfect (over 0.90).

4. Extrinsic evaluation: Advice game

To demonstrate the utility of implementing acoustic-prosodic entrainment in a spoken dialogue system, we created GoFishWithHelpers, a game in which subjects interact with an entraining avatar and a non-entraining one. GoFishWithHelpers is designed to test the hypothesis that speakers will trust an entraining avatar more than one that does not entrain. In addition, we explicitly ask for the subjects' impressions of each avatar, hypothesizing that they will prefer the entraining one and its voice.

4.1. GoFishWithHelpers

Figure 2 shows a screenshot of GoFishWithHelpers in progress.

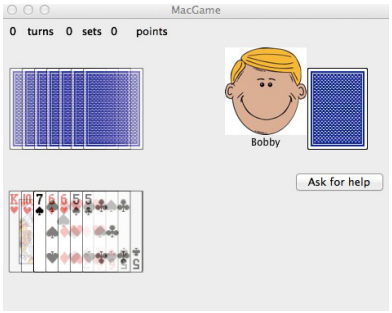

Figure 2: Screenshot from GoFishWithHelpers

In GoFishWithHelpers, trust is modeled by requiring the user to ask advice from one of two avatars at each turn. In every game, one avatar entrains to the subject while the other randomly changes its prosody. In the instructions before the game, subjects are told:

> *"Bobby and Freddy can see the system's hand and will tell you which rank to ask for. They will usually give you good advice, but sometimes they will give you bad advice. One will give you bad advice more often than the other. Your role in this game is to decide who to trust."*

Each helper is programmed to give advice according to an algorithm that keeps the persona's global advice score — the overall perceived quality of the advice it has given so far, corresponding with the number of points earned by following that advice — as close to zero as possible. At each turn, the helper asked for advice selects the rank whose score would bring the helper's global advice score closest to zero.

The game design objective of obscuring the *true* quality of an avatar's advice and of ensuring that each avatar gives a similar quality of advice are meant to prevent users from deciding whom to trust based on performance, and instead force her to rely subconsciously on paralinguistic cues. The player cannot alternate between helpers for the first few turns and then rely on the one that performed best on those turns for the remainder of the game — a strategy some subjects used in the pilot study — because the *perceived* quality of each helper's advice will be the same after several turns. However, the player is aware that *perceived* quality does not necessarily reflect *true* quality, which allows for the (false) illusion that one helper is in fact giving better advice.

The "Ask for help" button triggers recording. When the player has finished speaking, she toggles the same button (which now says "Done"), finalizing the recording and sending it to ASR and to the feature extractor. Once these are completed, the program identifies the name of the requested helper from the user input, and the advice output is formulated based on the helper's global advice score. In parallel, the entrainer takes the extracted features and uses them as input to determine the target prosody for the system's next utterance, based on the experimental condition (see Section 5). The target features are then added to the planned TTS input as SSML markup.

The selected helper's face appears on the screen and his voice provides advice (e.g. "Ask for nines"); the cards of the rank that the helper has chosen are enabled in the player's hand. The player completes the turn by clicking the enabled cards, which serves as the "request" for that rank from the system; the system either gives the player cards of that rank from its hand and awards the associated points or deals the player a card from the deck and deducts 50 points. At that point, the system is dealt a new hand from the deck and the turn is over. There are 15 turns in a game, and each subject plays three games.

4.2. Survey and advice score

After three games have been completed, the subject is directed to a short survey that asks the following questions:

- Which adviser did you like better?
- Whose voice did you like better?
- Who gave better advice?
- Please check the adjectives you would use to describe {Bobby's, Freddy's} (NB: Order of presentation of helper names in questions and choices is balanced across experimental conditions.) voice. – Loud – Soft – Pleasant – Annoying – Fast – Slow – Natural – Strange

The choice between Bobby and Freddy is forced for the first three questions. Subjects are also asked to complete TIPI [32], a short personality test measuring the Big Five personality traits.

The game is designed to implicitly model confidence in an avatar by requiring the player to decide whom to ask for advice at each turn; this choice is framed in the instructions by telling the player that "Your role in this game is to decide who to trust." Two confidence scores are calculated for each helper in a game. The **raw score** is the number of times the player asks a particular helper for help throughout the game. The **weighted score** is the sum of the indices of turns in which that helper is asked for advice, with the intuition that trust in later turns should be weighted more heavily as an indicator of developing trust. The subjects were also explicitly asked in the post-study survey to choose which helper they think gave better advice.

5. Pilot studies

Three pilot studies were carried out to evaluate the prosodically entraining algorithm using the advice game described in Section 4: one in New York with speakers of Standard American English, one in Buenos Aires with speakers of Argentine Spanish, and one in Bratislava with Slovak speakers. The three differ slightly in implementation and findings, but together indicate that the approach is feasible and that the association between confidence in an avatar and its acoustic-prosodic entrainment behavior is a promising research direction.

English The English pilot study used PocketSphinx [33] for automatic speech recognition (ASR), and Cepstral 6, a unit-selection-based TTS. The two avatars, Bobby and Freddy, each had a different male voice. In each trial, one avatar entrained on speech rate and intensity, and the other sampled random rate and intensity values from a range of values previously determined to sound natural. Nineteen subjects participated in the study. A linear regression with trust score as the dependent variable and entrainment condition (entrain vs. random) and avatar persona (Bobby vs. Freddy) found that the no-entrainment control was a significant predictor of a lower trust score ($p < 0.001$). The post-interaction survey revealed that subjects liked the entraining voice better ($\chi^2(1) = 3.74, p = 0.053$) and were less likely to call it "annoying" ($\chi^2(1) = 9.32, p = 0.0023$).

Spanish The Spanish implementation also used PocketSphinx [33] for ASR, and a MaryTTS HMM voice built with a corpus read by a female professional speaker of Argentine Spanish for TTS [34]. When asking for advice, subjects chose between two female avatars, Amanda and Eugenia, the differ-

Language	Avatar gender	Entrainment		Baseline	Entrainment × Advice score
		Features	Method		
English	Male	Intensity Speech rate	Absolute	Random	+ ($p < 0.001$)
Spanish	Female	Speech rate	Relative	Disentrain	- ($p < 0.1$)
				Constant	no effect
Spanish	Female	Speech rate	Relative	Disentrain	- ($p < 0.05$)
				Constant	no effect

Table 1: Summary of pilot experiments

ent voices were simulated by using two different pitch ranges. The only prosodic feature entrained on was speaking rate, measured in syllables per second from ASR output. In condition 1, one avatar entrained (matching the subject's relative changes in speech rate), and the other disentrained (moved in the opposite direction). In condition 2, one avatar entrained, while the other always used the default speech rate. 14 subjects participated in condition 1, and 12 in condition 2. The advice score (number of times an avatar was asked for advice) was regressed against the entrainment condition (entrains vs. disentrains in condition 1, entrains vs. default in condition 2). Avatar identity was included as an independent variable to control for possible effects. Results show that, in condition 1, subjects were slightly *less* likely to ask the entraining avatar for advice (approaching significance at $p < 0.1$). In condition 2, the entrainment behavior had no significant effect. The avatar identity had no significant effect in either condition.

Slovak Slovak ASR used the open-source decoder Julius (http://julius.osdn.jp/en_index.php) complemented by triphone acoustic models and a Slovak language model developed at the Institute of Informatics, Slovak Academy of Sciences. Slovak TTS is based on statistical parametric synthesis using the open-source engine HTS (http://hts.sp.nitech.ac.jp) with adjustments to f0, intensity, and speech rate built by II SAS from a corpus of sentences read by a Slovak female professional speaker. The two avatars were named Monika and Tereza and were created using a single female voice. The experimental design paralleled that of the Spanish pilot, with 10 subjects in each condition (entrain vs. disentrain, entrain vs. default). As in the Spanish pilot, the only feature entrained on was speech rate. In condition 1, the linear regression revealed a *negative* relationship between entraining on speech rate and advice score ($p < 0.05$). In condition 2, entrainment behavior had no significant effect. These results parallel the findings of the Spanish pilot, with a slightly stronger effect of entrainment in condition 1.

5.1. Discussion

The experimental designs and results of the pilot studies are summarized in Table 1. In the English pilot, entrainment had the predicted association with advice score, with subjects more likely to ask advice from the entraining avatar. The Spanish and Slovak studies, however, found a preference for disentraining avatars. This discrepancy may stem from any number of the differences in implementations between the studies, from the entrainment features (only the English study entrained on speaking rate), the baseline (both between and within studies, the baseline had an effect), to the avatar persona (the English avatars were male; the Spanish and Slovak avatars were female). In a study of the interactions between entrainment and dialogue quality in human-human conversations, both gender and the entrainment feature(s) were shown to have an effect [29]. The consistency in results between the Spanish and Slovak studies, which had identical designs, validates this extrinsic evaluation and suggests that future work can more systematically vary entrainment variables and control conditions to further explicate the effect of an avatar's entrainment behavior.

6. Conclusion

This work presents an initial architecture for incorporating knowledge of acoustic-prosodic entrainment into a virtual conversational agent: observational studies here become blueprints for a system design. Furthermore, an entraining system can be a valuable mechanism for testing associations between entrainment and dialogue quality in the absence of confounding factors. With respect to SDS, this work introduces new possibilities for improving the user experience in a way that is orthogonal to improvements that can be made to a system's core components. The design of this method is informed by our research on entrainment in human-human conversations, but does not fully implement the range of entrainment behaviors observed in humans. Future work will address the task of translating the accumulated empirical observations of human-human entrainment into algorithms suitable for SDS. Another direction for future work is the exploration of alternative methods for implementing the adaptation step, such as HMM-based voice conversion (e.g. [35]). Such methods may have improved flexibility and sound more natural, but they sacrifice the encapsulation of the approach described here, which can interface with any TTS.

We validate the entrainment method in an experimental context with three pilot experiments in English, Slovak and Spanish. The English study showed a positive association between entrainment and an avatar's perceived reliability and likability. The Spanish and Slovak studies, with a different experimental design, showed a negative association in one condition. This discrepancy is in line with work on human-human conversations, which shows that the interaction between entrainment and dialogue quality differs with the speaker gender and the entrained features. Encouragingly, the Spanish and Slovak results, which had identical experimental designs, were consistent with each other.table This work also confirms experimentally that current TTS technologies are capable of conforming to specific prosodic feature values for intensity, pitch, and speaking rate, using both HMM and concatenative synthesis. Together, the contributions of this work set the stage for future work in the space of acoustic-prosodic entrainment by a virtual conversational agent.

7. Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-15-1-0055.

8. References

[1] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proceedings of Interspeech*, 2010.

[2] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011.

[3] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, 1999.

[4] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[5] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1994, pp. 72–78.

[6] R. Levitan, "Acoustic-prosodic entrainment in human-human and human-computer dialogue," Ph.D. dissertation, Columbia University, 2014.

[7] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability," *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975.

[8] S. Gregory, S. Webster, and G. Huang, "Voice pitch and amplitude convergence as a metric of quality in dyadic interviews," *Language & Communication*, vol. 13, no. 3, pp. 195–217, 1993.

[9] A. Ward and D. Litman, "Measuring convergence and priming in tutorial dialog," University of Pittsburgh, Tech. Rep., 2007.

[10] S. E. Brennan, "Lexical entrainment in spontaneous dialog," *Proceedings of ISSD*, pp. 41–44, 1996.

[11] H. P. Branigan, M. J. Pickering, and A. A. Cleland, "Syntactic coordination in dialogue," *Cognition*, vol. 75, no. 2, pp. B13–B25, 2000.

[12] D. Reitter, J. D. Moore, and F. Keller, "Priming of syntactic rules in task-oriented dialogue and spontaneous conversation," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006, p. 685690.

[13] K. G. Niederhoffer and J. W. Pennebaker, "Linguistic style matching in social interaction," *Journal of Language and Social Psychology*, vol. 21, no. 4, pp. 337–360, 2002.

[14] C. Danescu-Niculescu-Mizil, M. Gamon, and S. Dumais, "Mark my words! linguistic style accommodation in social media," in *Proceedings of WWW*, 2011.

[15] L. Michael and J. Otterbacher, "Write like I write: Herding in the language of online reviews," in *Proceedings of the Eigth International AAAI Conference on Weblogs and Social Media*, 2014.

[16] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context, and consequence," *Contexts of accommodation: Developments in applied sociolinguistics*, vol. 1, 1991.

[17] R. Y. Bourhis and H. Giles, "The language of intergroup distinctiveness," *Language, ethnicity and intergroup relations*, vol. 13, p. 119, 1977.

[18] R. L. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[19] D. Reitter and J. D. Moore, "Predicting success in dialogue," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 808–815.

[20] M. E. Ireland, R. B. Slatcher, P. W. Eastwick, L. E. Scissors, E. J. Finkel, and J. W. Pennebaker, "Language style matching predicts relationship initiation and stability," *Psychological Science*, vol. 22, no. 1, pp. 39–44, 2011.

[21] J. Thomason, H. V. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *Artificial Intelligence in Education*. Springer, 2013, pp. 750–753.

[22] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2015.

[23] Z. Hu, G. Halberg, C. R. Jimenez, and M. A. Walker, "Entrainment in pedestrian direction giving: How many kinds of entrainment?" in *Proceedings of 5th International Workshop on Spoken Dialog System*, 2014.

[24] J. Lopes, M. Eskenazi, and I. Trancoso, "Automated two-way entrainment to improve spoken dialog system performance," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8372–8376.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2012, version 5.3.23, retrieved 21 August 2012 from http://www.praat.org.

[26] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Speech Prosody*, 2004.

[27] A. Rosenberg, "AuToBI - a tool for automatic ToBI annotation," in *Proceedings of Interspeech*, 2010, pp. 146–149.

[28] R. Levitan, Š. Beňuš, A. Gravano, and J. Hirschberg, "Entrainment in slovak, spanish, english, and chinese: A cross-linguistic comparison," in *Proceedings of SIGdial*, 2015.

[29] R. Levitan, A. Gravano, L. Willson, S. Benus, 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*. Montréal, Canada: Association for Computational Linguistics, June 2012, pp. 11–19. [Online]. Available: http://www.aclweb.org/anthology/N12-1002

[30] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *AAAI 2015 Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[31] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system MARY: A tool for research, development, and teaching," in *SSW*, 2001.

[32] S. Gosling, P. Rentfrow, and W. Swann, "A very brief measure of the big-five personality domains," *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, 2003.

[33] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006.

[34] L. Violante, P. R. Zivic, and A. Gravano, "Improving speech synthesis quality by reducing pitch peaks in the source recordings." in *HLT-NAACL*, 2013, pp. 502–506.

[35] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-hmms for expressive speech synthesis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1109–1116, 2006.