The effect of soft, modal and loud voice levels on entrainment in noisy conditions

Éva Székely, Mark T. Keane, Julie Carson-Berndsen

School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
eva.szekely@ucdconnect.ie, mark.keane@ucd.ie, julie.berndsen@ucd.ie

Abstract

Conversation partners have a tendency to adapt their vocal intensity to each other and to other social and environmental factors. A socially adequate vocal intensity level by a speech synthesiser that goes beyond mere volume adjustment is highly desirable for a rewarding and successful human-machine or machine mediated human-human interaction. This paper examines the interaction of the Lombard effect and speaker entrainment in a controlled experiment conducted with a confederate interlocutor. The interlocutor was asked to maintain either a soft, a modal or a loud voice level during the dialogues. Through half of the trials, subjects were exposed to a cocktail party noise through headphones. The analytical results suggest that both the background noise and the interlocutor's voice level affect the dynamics of speaker entrainment. Speakers appear to still entrain to the voice level of their interlocutor in noisy conditions, though to a lesser extent, as strategies of ensuring intelligibility affect voice levels as well. These findings could be leveraged in spoken dialogue systems and speech generating devices to help choose a vocal effort level for the synthetic voice that is both intelligible and socially suited to a specific interaction.

Index Terms: entrainment, vocal intensity, Lombard effect, adaptive interaction

1. Introduction

Speakers continuously adjust their speech delivery in response to context, using a variety of strategies to aid the conversation flow and to ensure that their speech is intelligible. This aspect of inter-personal communication poses extra challenges for people who use Augmentative and Alternative Communication (AAC). Besides flexibility and variety of expression, context awareness is needed in Speech Generating Devices (SGDs) to respond to the physical and social environment [1]. To support conversational interaction for people who rely on communication devices, the automatic adjustment of vocal intensity levels as well as the active involvement of the communication partner to suit differing communicator roles and relationships is desirable [2]. Spoken dialogue systems and virtual conversational agents would greatly benefit from such context dependent flexibility and the ability to take into account both environmental and social aspects of conversation [3, 4, 5].

One of the ways people adapt their speech to the environment is by increasing their vocal intensity if there is background noise. This is referred to as the *Lombard effect* [6, 7]. While the Lombard effect has originally been viewed as an automatic regulation of voice intensity based on auditory feedback, it has been shown that a greater level of communicative involvement, such as being immersed in a live conversation with an interlocutor (as opposed to reading out loud to a recording device), increases the Lombard effect. In experiments conducted by [8], the presence of a conversation partner increased the vocal intensity modification in subjects' speech, and it also elicited other speaker dependent communicative strategies.

Apart from increased vocal intensity, speech produced in noise also shows characteristics of decreased speech rate, a spectral shift of energy towards the medium frequencies, increased tension of the vocal folds, more pronounced articulation and phoneme modification [9]. The distinctive features of Lombard speech and of differing levels of vocal effort have been widely studied [9, 10], and integrated into synthetic speech [11, 12]. We have therefore now the prospect of integrating context sensitivity and flexibility in vocal intensity levels into a system much more proficiently than merely automatising the volume adjustment of the Text-to-Speech engine.

Another phenomenon characterising the way speakers adapt to context is the tendency of conversation partners to behave similarly by adapting to each others' speech. This phenomenon is called entrainment (also referred to as alignment, accommodation or adaptation) [13], and takes place in many acoustic, phonetic, prosodic, syntactic and lexical dimensions [14, 15, 16]. Entrainment has been shown to contribute to the naturalness and success of a conversation and the speakers' degree of involvement in a conversation [4, 17].

Entrainment also occurs in human-machine dialogues [18, 3]. Vocal intensity, in particular, has proven to be a feature that elicits adaptation in response to synthetic speech [19]. Therefore, for a machine to be able to choose a suitable vocal intensity level for a conversation, strategies of active listening are desirable, that go beyond responding to the level of the background noise. It is also not sufficient however, to simply mimic the human conversation partner's vocal behaviour because differing levels of background noise can influence intelligibility and subsequently may also affect the dynamics of speaker entrainment during the course of an interaction.

There are also many other aspects of context that influence the vocal intensity level conversation partners choose during an interaction. The physical distance between the speakers [20], the social setting [21], the speaker's affect cueing [22], attitude [23], personality [24] and the relationship between the conversation partners [9], and inevitably, the topic of the dialogue can greatly impact how loudly a conversation takes place. As an example, when sharing about a confidential subject, or if there are other persons in the room, a speaker may limit the measure with which they increase their vocal intensity in response to background noise. At the same time, a speaker may raise

their voice level in silence to account for a greater physical distance between them and their interlocutor, or to express their dominant position in the social situation.

It would be very difficult for a system to take all of these factors into account, but it is safe to assume, that the human partner in the human-machine (or machine mediated human-human) dialogue does take them into consideration. If this is true, then as long as the machine is capable of computing a socially appropriate *response* to a particular vocal intensity level in a particular noise level, the flow of the conversation may resemble a more natural interaction. In this paper, we designed a controlled experiment that aimed to make socially appropriate estimations for system responses to a human conversation partner.

2. The Experiment

The experiment aimed to simulate a human-machine interaction situation, where a confederate interlocutor, henceforth referred to as the *interlocutor*, speaking with different levels of vocal effort, played the role of the human speaker. Subjects were recruited to play the role of the machine interlocutor, so that a machine's desired behaviour could be derived from the subjects' vocal behaviour in the experiment. Using a human interlocutor here was necessary to display a natural array of vocal intensity levels, as the mere amplitude modification of pre-recorded utterances does not appear to elicit speaker adaptation when speakers are exposed to background noise [25]. The subjects were asked to play an interactive card matching game with the interlocutor which required a high level of cooperation since the matches were not exact but based on topical similarity. This resulted in task oriented dialogues, on average 3.5 minutes in length. The experiment consisted of a 2 x 4 within-subject design that crossed Background Noise (Silence or Noisy) and Voice Levels (*Reference*, *Soft*, *Modal* or *Loud*). The main manipulation of the Voice Level variable was implemented by the interlocutor maintaining a soft, modal or loud voice. In addition, the Reference Condition was used as a check, in which the interlocutor was asked not to be specifically conscious of his voice and focus on the game instead. The voice levels realised by the interlocutor can be described along the vocal effort continuum, where *Soft* Voice is decreased vocal effort with voicing (so not whispering) and *Loud* Voice is increased vocal effort level that can be placed between *Modal* Voice and shouting.

The experimental conditions are summarised in Table 1:

Interlocutor's voice	Silence	Noise
Reference	CO1	CO2
Soft	CO3	CO6
Modal	CO4	CO7
Loud	CO5	CO8

Table 1: Experimental Conditions

The noise that the subjects heard (cocktail party noise at 90db) was played through closed headphones, which was necessary to avoid interfering with the interlocutor's effort to maintain a specific voice level in each condition. The subject's own voice was played back into the headphones to provide additional self-monitoring feedback. This has been shown to have a slight compensating effect aiding the Lombard effect taking place similar to when the noise is played through loudspeakers [8]. A short break was held between the sessions, to limit the interlocutor's vocal fatigue, and a screen was placed in between the participants to ensure that all communication remained verbal.

Twelve volunteer subjects, 10 males and 2 females aged between 18 and 54 (averaging 27) participated in the experiments. The role of the confederate interlocutor was played by a 28 year old male native English speaker. The recordings were conducted using Behringer Super Cardioid XM1800S microphones. The headphones used were Sony Stereo MDR XD100's. All audio was recorded on a 16 gigabyte SD Card using a Zoom H4n Handy Mobile Recorder. The recordings were sampled at 44.1 kHz (16 bits, mono), to eliminate the effect of clipping; then downsampled to 16 kHz.

After the sessions were finished, the subjects completed a questionnaire in which they were asked what they thought the experiment was measuring. The questionnaire revealed that only 2 out of the 12 participants suspected that the experiment had anything to do with their speech, and none of the 12 participants suspected that their conversation partner was a confederate, but rather another participant in the experiment. This means that the experiment succeeded in creating a setting where the semi-conscious phenomena of speaker accommodation and Lombard effect could be studied in different conditions of a task-oriented dialogue.

3. Measuring Entrainment

3.1. Speech Features

The focus of the experiment is to show how the interlocutor's different voice levels and the background noise together affected the subjects' vocal intensity. The following acoustic features were selected to characterise vocal intensity:

perceived loudness To estimate the perceived vocal intensity of the speakers, an approximate measure of perceived loudness was used: normalised intensity raised to the power of 0.3 simulate human sensitivity to loudness. This feature was extracted using openSMILE [26].

fundamental frequency *f0*, as extracted with openSMILE [26] via the Sub-Harmonic-Summation (SHS) method.

voice quality A wavelet-transform based voice quality feature, characterising voice qualities on a breathy to tense dimension called Peak Slope [27] was extracted.

3.2. Significant Difference in Conditions

Before we can look at how subjects adapted to the interlocutor's speech in the different experimental conditions, we need to validate that the interlocutor indeed maintained a different voice level corresponding to *Soft*, *Modal* and *Loud* vocal intensity in each of the three conditions. We do this by computing the difference between the mean loudness feature of the interlocutor's speech over whole sessions for the three conditions, looking for significance, and for the *f0* and Peak Slope features for further characterisation of these voice levels.

3.3. Global Similarity

Entrainment on a global level occurs if a particular feature of a speaker's speech, is similar to that of her interlocutor over a whole conversational segment. This can be measured by calculating feature means over an entire session and comparing the difference between conversational partners with the differences between non-partners in the same corpus, or by comparing a speaker to herself, talking to a different interlocutor, or to the

same interlocutor but among different conditions [3]. In this experiment we have access to data where the speaker and the communication setting is identical, with the confederate interlocutor's voluntary voice regulation being the only factor defining a different condition. In this setting, entrainment of the subject to the controlled interlocutor voice can be evaluated based on the distance between subjects' speech and the interlocutor's speech in a session, versus the distance between the subject's speech in the same session and the interlocutor's speech in the two other sessions within the same environmental condition, as shown in equations 1 and 2 below.

$$ENT(s,c) = -\mid f_s^c - f_i^c \mid \quad (1)$$

$$ENTX(s,c) = -\frac{\sum_{x<>c} \mid f_s^c - f_i^x \mid}{n_c - 1} \quad (2)$$

where s : subject, c : condition, f : speech feature measured, i : interlocutor, n_c : number of conditions

3.4. Local Entrainment

Local entrainment encompasses different ways in which dynamic alignment occurs between interlocutors within a conversation. Global and local entrainment operate independently within a dialogue. Local entrainment is therefore to be calculated at the turn level rather than at the session level. [14] defines turn as a maximal sequence of inter-pausal units from a single speaker. Inter-pausal units (IPUs) refer to pause-free speech sequences from one speaker, separated from one another by at least 50ms. In our calculations, following the method presented in [14] the last IPU of each turn was compared with the first IPU of the subsequent turn.

3.4.1. Synchrony

Synchrony is a form of local entrainment, which is concerned with the relationship between interlocutors' relative values. Positive synchrony can be observed when speakers simultaneously show similar behaviours on the turn level, meaning that if one speaker raises their vocal intensity, their interlocutor will also raise their vocal intensity in the next turn [28]. Synchrony in the negative direction, referred to as *complementary entrainment* or *asynchrony* [4], is realised when the speakers mimic an aspect of each others' vocal behaviour, but in the opposite direction. For vocal intensity, this means that if one interlocutor speaks more softly in a turn, that will be met by a louder utterance by the other speaker in the next turn.

3.4.2. Convergent versus Divergent Entrainment

Convergence is an aspect of local entrainment where the similarity between the two speakers increases over time; their speech accommodating towards a common point [28]. If the two speakers move apart in different directions during the course of the conversation, this is referred to as *divergence*.

Convergence is measured by the slope of a linear regression over time (x-axis) and the difference in vocal intensity between the interlocutor and subject in the following utterance (y-axis). As the similarity measure itself is always negative, a positive slope over time indicates convergent entrainment. Under the specific conditions of our experiment, the subject is expected to exhibit the majority of the adaptation in vocal intensity, but the direction and speed of the entrainment is still a marker of common communicative adaptive behaviour.

4. Results

4.1. Difference between Conditions

For the mean of the loudness parameter over the interlocutor's utterances, we found that the intensity of his voice was significantly different for each of the different voice level Conditions in both the Silent and the Noisy Condition ($p < 0.001$*).

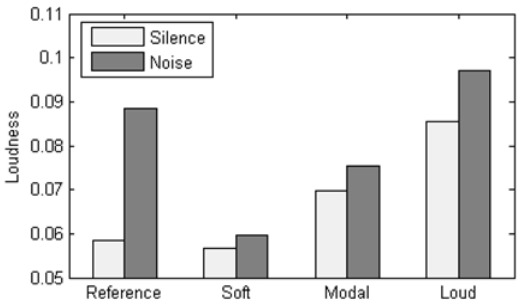

Figure 1: Average loudness of the interlocutor's speech in the different conditions

As shown in Figure 1, the interlocutor's voice was slightly louder in all cases when the subjects heard a background noise indicating a reverse direction entrainment in response to the elevated vocal intensity of his conversation partners caused by the Lombard effect. In the Reference Condition, this effect is very significant, with the interlocutor's voice in the Noisy environment Reference Condition being 51.8% louder than in the Silent environment Condition. However, this difference is not significant in the conditions where he is voluntarily controlling his voice to reflect soft, modal and loud levels.

	soft	modal	loud	p-value
loudness	0.058	0.073	0.091	<0.001*
mean *f0*	85.5	89.2	96.0	<0.001*
PeakSlope	-0.318	-0.353	-0.379	<0.001*

Table 2: Feature levels of the interlocutor's speech and outcome of one-way repeated measures ANOVA for each feature

Table 2 summarises the average values for each voice level of the interlocutor. His vocal intensity increase as shown by the loudness parameter, appears to correlate to a significant *f0* increase. The PeakSlope parameter shows a steeper decline as vocal intensity increases, indicating the three voice levels' place on a lax-to-tense continuum. This is consistent with the findings of [23] and [22] showing that decreased vocal intensity can be associated with characteristics of breathy phonation while the increased vocal effort in louder voice levels contributes to increased tension of the vocal folds.

While there is variance within all sessions that causes some overlap when looking at the mean features of individual utterances, the distribution of these features are significantly different for each voice level. Hence we conclude that the interlocutor succeeded in maintaining three distinct voice levels that can be categorised as *Soft, Modal* and *Loud* Voice throughout the trials, defining the different controlled Voice Level conditions.

In examining the subject's uncontrolled speech, the Lombard effect was clearly measurable, with the subject's average loudness value being 0.077 in the Silent Condition, and 0.100 in the Noisy Condition. For the remainder of this section we will focus on showing the results of entrainment measures between interlocutor and subjects in the *loudness* parameter.

4.2. Global similarity

Entrainment on a global level is defined by a smaller distance between the parameters of subject and interlocutor within the same conversation than between a subject and interlocutor that are *not* in the same conversation (and hence under different conditions). These distances are defined as negatives, hence convergent entrainment is found if $ENT(s,c) - ENTX(s,c) > 0$.

Under the six conditions, entrainment is significant for the Modal and Loud Voice Level in Silence, and for the Loud Voice Level in Noise (see Table 3).

	Silence		Noise	
	t-value	p-value	t-value	p-value
Soft	-0.647	0.735	-4.315	0.999
Modal	3.597	0.002*	0.042	0.484
Loud	1.933	0.040*	4.781	<0.001*

Table 3: Global similarity: T-tests for global entrainment in different voice level and environment conditions

4.3. Local entrainment of subjects and interlocutor

4.3.1. Synchrony

Synchrony between interlocutor (x-axis) and subject (y-axis) is measured by the slope of a linear regression on the paired changes in vocal intensity, as the subject changes his vocal intensity following a change by the interlocutor. Table 4 shows the results of the synchrony measure.

Positive synchrony was found where the interlocutor spoke with a Loud Voice in both Silent and Noisy Conditions and Modal Voice in the Noisy Condition. Strong complementary entrainment (synchronous change in the opposite direction) occurred where the interlocutor applied a Soft Voice in a Noisy Condition.

	Reference	Soft	Modal	Loud
Silence	0.13	-0.01	0.02	0.24
Noise	0.15	-0.86	0.24	0.41

Table 4: Synchrony: average similarity between speakers in loudness change at turn exchanges

Figure 2 shows examples of negative synchrony in the Soft Voice Level & Noisy Condition and positive synchrony in the Loud Voice Level & Noisy Condition.

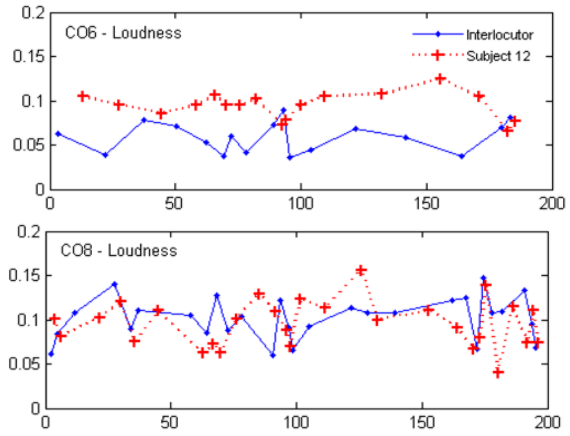

Figure 2: Time series of IPUs of subject 12 (red) and the interlocutor (blue) in the Soft Voice Level & Noisy Condition (CO6) and in the Loud Voice Level & Noisy Condition (CO8).

4.3.2. Convergence

A positive convergence slope indicates a decreasing difference in loudness between the speakers over the course of the conversations. The interlocutor's choice of controlled voice level did not appear to have a significant effect on the convergence in the dialogues. However, when looking at all conversations in the Silent and the Noisy Condition, significant convergence is measured in the Silent Condition, while convergence in the Noisy Condition approaches significance, as shown in Table 5.

	Slope estimate	t-value	p-value
Silence	0.0056	2.369	0.018*
Noise	0.0068	1.937	0.053

Table 5: Convergence: T-tests for convergence over time in loudness between subjects and interlocutor

5. Discussion

While the presence of background noise elicited a significantly louder vocal intensity in the subjects' speech in all conditions, global entrainment was found to be most significant when the interlocutor was speaking with a loud vocal intensity. When talking to someone who speaks softly in a noisy environment, speakers appear to adapt a strategy of raising their own vocal intensity in response in the next turn, as demonstrated in the complementary entrainment of the synchrony measure. This can be interpreted as an active listening strategy to encourage the other person to speak louder without having to explicitly prompt them to do so. The presence of background noise attenuated, but did not fully eliminate convergence in conversations.

The results of this work could be used in an adaptive dialogue system or SGD in at least two ways. In a purely data-driven approach, vocal intensity response models could be trained on the recordings for the system to classify the situation based measuring the background noise together with the conversation partners voice and compute a suitable intensity value for the next utterance. Additionally, these findings can be useful in making design decisions, for example, on how many turns to take into account for a decision, or in which situations to switch to a different strategy of modelling entrainment.

6. Conclusion and Future Work

Both the environmental conditions and the interlocutor's controlled voice levels had a significant effect on what intensity level subjects chose in a dialogue. This effect was observed on the global level, in conversations as a whole as well as on the turn level, influencing the dynamics of entrainment between speakers. Based on these results we can conclude that if a system is to be comprehensively context sensitive, it would have to consider the level of background noise together with the conversation partner's voice level to compute an appropriate vocal intensity for an utterance. Future work involves studying further dimensions of speaker entrainment on this data, such as adaptation on the lexical and temporal domain.

7. Acknowledgements

This research is partly supported by Science Foundation Ireland through the CNGL Programme (Grant 12/CE/I2267) in the ADAPT Centre (www.adaptcentre.ie) at University College Dublin. The authors wish to thank Peter Ward, Rohan Jain and Céline De Looze for their assistance.

8. References

[1] J. L. Arnott and N. Alm, "Towards the improvement of augmentative and alternative communication through the modelling of conversation," *Computer Speech & Language*, vol. 27, no. 6, pp. 1194–1211, 2013.

[2] D. J. Higginbotham, H. Shane, S. Russell, and K. Caves, "Access to aac: Present, past, and future," *Augmentative and Alternative Communication*, vol. 23, no. 3, pp. 243–257, 2007.

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

[4] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, pp. 11–34, 2014.

[5] A. Hönemann and P. Wagner, "Adaptive speech synthesis in a cognitive robotic service apartment: An overview and first steps towards voice selection," *Elektronische Sprachsignalverarbeitung 2015*, 2015.

[6] E. Lombard, "Le signe d'élévation de la voix [the sign of the elevation of the voice]," *Annales des maladies de l'oreille et du larynx*, vol. 37, pp. 101–119, 1911.

[7] W. Van Summers, D. B. Pisoni, R. H. Bernacki, R. I. Pedlow, and M. A. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *The Journal of the Acoustical Society of America*, vol. 84, no. 3, pp. 917–928, 1988.

[8] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the lombard effect," *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 3, pp. 588–608, 2010.

[9] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543–571, 2014.

[10] M. Charfuelan and M. Schröder, "The vocal effort of dominance in scenario meetings." in *Proceedings of Interspeech*, 2011, pp. 2953–2956.

[11] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[12] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[13] M. J. Pickering and S. Garrod, "Alignment as the basis for successful communication," *Research on Language and Computation*, vol. 4, no. 2-3, pp. 203–228, 2006.

[14] R. Levitan and J. B. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech*, 2011, pp. 3081–3084.

[15] 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.

[16] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 6, p. 1482, 1996.

[17] Š. Beňuš, "Social aspects of entrainment in spoken interaction," *Cognitive Computation*, vol. 6, no. 4, pp. 802–813, 2014.

[18] A. Fandrianto and M. Eskenazi, "Prosodic entrainment in an information-driven dialog system." in *Proceedings of Interspeech*, 2012, pp. 342–345.

[19] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in childrens conversational speech with animated personas," in *Proceedings of the 7th International Conference on Spoken Language Processing*, vol. 4, 2002, pp. 2689–2692.

[20] H. A. Cheyne, K. Kalgaonkar, M. Clements, and P. Zurek, "Talker-to-listener distance effects on speech production and perception," *The Journal of the Acoustical Society of America*, vol. 126, no. 4, pp. 2052–2060, 2009.

[21] 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*. Association for Computational Linguistics, 2012, pp. 11–19.

[22] I. Yanushevskaya, C. Gobl, and A. N. Chasaide, "Voice quality in affect cueing: does loudness matter?" *Frontiers in psychology*, vol. 4, 2013.

[23] I. Carlos Toshinori, I. Hiroshi, and H. Norihiro, "Analysis of the roles and the dynamics of breathy and whispery voice qualities in dialogue speech," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[24] M. Charfuelan, M. Schröder, and I. Steiner, "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings." in *Proceedings of Interspeech*, 2010, pp. 2558–2561.

[25] R. S. Tweedy and J. F. Culling, "Does the signal-to-noise ratio of an interlocutor influence a speaker's vocal intensity?" *Computer Speech & Language*, vol. 28, no. 2, pp. 572–579, 2014.

[26] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[27] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform." in *Proceedings of Interspeech*, 2011, pp. 177–180.

[28] J. Edlund, M. Heldner, and J. Hirschberg, "Pause and gap length in face-to-face interaction," in *Proceedings of Interspeech*, 2009, pp. 2779–2782.