Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Prosody and hand gesture at turn boundaries in Swedish

Margaret Zellers[1], *David House*[2], *Simon Alexanderson*[2]

[1]Department of Linguistics: English, University of Stuttgart, Germany
[2]Department of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

margaret.zellers@ifla.uni-stuttgart.de, davidh@speech.kth.se, simonal@kth.se

Abstract

In order to ensure smooth turn-taking between conversational participants, interlocutors must have ways of providing information to one another about whether they have finished speaking or intend to continue. The current work investigates Swedish speakers' use of hand gestures in conjunction with turn change or turn hold in unrestricted, spontaneous speech. As has been reported by other researchers, we find that speakers' gestures end before the end of speech in cases of turn change, while they may extend well beyond the end of a given speech chunk in the case of turn hold. We investigate the degree to which prosodic cues and gesture cues to turn transition in Swedish face-to-face conversation are complementary versus functioning additively. The co-occurrence of acoustic prosodic features and gesture at potential turn boundaries gives strong support for considering hand gestures as part of the prosodic system, particularly in the context of discourse-level information such as maintaining smooth turn transition.

Index Terms: gesture, turn transition, multimodal communication, Swedish

1. Introduction

When interlocutors converse, they tend to do so with a minimum of problematic overlaps or large silent gaps ([1],[2]). In order for this to be the case, speakers and listeners must have ways of indicating to one another that they intend to speak or to stop speaking.

Studies on turn-taking from a variety of theoretical viewpoints have identified a number of linguistic features that play a role in signaling turn transition, including syntactic/semantic completion (e.g. [3]; [4], [5]), intonational features (e.g. [6]; [7]; [8]), and phonation quality/spectral characteristics (e.g. [9]; [10]). Unsurprisingly, several recent studies (e.g. [11]; [12], [13]; [14]) report a hierarchy of the various features correlated with turn transition or turn hold, including both lexicosyntactic as well as phonetic features.

The prosodic cues which have priority when it comes to signaling turn transition appear to vary from language to language. Many English varieties are reported to make heavy use of intonational cues (cf. [12], [13]), while a perception study showed that listeners apparently do not attend to variations in segment duration ([15]). In Central Swedish, however, perception of turn change or hold appears to be more dependent upon the extent of final lengthening in syllables preceding potential turn boundaries ([16], [17]), although variation in final pitch has also been reported to be present ([14], [18], [19]).

In addition to linguistic cues, gesture has also begun to be considered as part of the turn-taking system. [20] and [21] report that conversational participants modulate their turn-taking behavior on the basis of the eye gaze behavior of a computer-generated avatar. [22] reports that conversational participants can use hand gestures to indicate their readiness to hear or invite an interlocutor to do the same. [23] report that gestures at transition relevance places may be used to project something to come; similarly, [24] and [25] report that gestures may be held until a shared understanding has been reached, which may involve inviting an interlocutor to demonstrate understanding by taking up a turn. [26] find that hand gestures are correlated with prosodic phrase boundaries and may thus contribute to the segmentation of speech into phrases.

The current study investigates the role that prosodic and gestural cues may play in combination with one another to help interlocutors achieve smooth turn transition. We consider the possibility that prosody and gesture may be similar linguistic systems or even aspects of the same system (cf. also [27]). The current study focuses specifically on hand gestures accompanying Swedish conversational speech.

2. Methodology

2.1. Data

Our data come from three five-minute segments of spontaneous conversations in Swedish drawn from the Spontal corpus ([28]). The conversations are conducted between two parties who are sitting face-to-face; our data include two male-male pairs and one male-female pair. Spontal includes high quality video, audio, and motion capture data, allowing for rich investigation of multimodal aspects of conversation.

2.2. Annotation

We extracted information about the acoustics and gesture features through a partially automated process. Talk-spurt ends were automatically detected by determining the speech activity at each frame. For this purpose, a voice activity detection (VAD) algorithm ([29]) was applied to the audio recordings from the near-field microphones attached to the subjects. The process resulted in two sets of speech/non-speech segments for each dialogue, one for each subject. Then, the final 500ms of speech preceding the ends of the talk-spurts were phonetically segmented by hand, and the talk-spurt boundaries were labeled according to whether speech was syntactically complete at that point. If the speech was syntactically complete, the boundary locations were further annotated on the basis of what happened next in the interaction, giving us four turn transition types as follows:

- **Backchannel:** The other speaker produced a backchannel or response token ([30]) following the talk spurt boundary, but did not take up a full turn

doi: 10.21437/SpeechProsody.2016-170

- **Change:** The other speaker took up a full turn following the boundary
- **Hold:** The same speaker continued speaking following the talk spurt boundary
- **Question:** The speech preceding the boundary was syntactically in the form of a question, and the other speaker took up the turn following the boundary

In addition to the segmental information, pitch information for the final pitch movement in each turn was extracted (when possible; many turns ended in creaky voice and were thus unusable for the pitch analysis). The current analysis uses the fundamental frequency (F0) of the final voiced frame in each 500ms window as the end pitch, and a hand-identified preceding F0 peak when available (note that pitch rises are relatively uncommon in Swedish at phrase ends (cf. [31], [32], so we include only pitch falls in the current analysis).

Hand gestures were identified through analysis of the video using ELAN ([33], [34]). Either one-handed or two-handed gestures were eligible for inclusion in the analysis, and are not separated in the current study. Gestures were segmented into their component parts on the basis of silent watching of the video, so the annotators did not have direct access to information about syntactic completion or turn transition. These component parts include the gesture phases of preparation, stroke, retraction, and pre- and post-stroke holds, following [35] and [36]. The stroke is the (generally dynamic) gesture phase which must be present in order for a gesture to have been said to have occurred; it involves a purposeful movement in the gesture space. The preparation and retraction are optional dynamic phases in which the hands are moved from rest to the gesture space, or return from the gesture space to rest, respectively. Finally, pre- or post-stroke holds are optional static phases preceding or following the stroke phase; that is, the hand(s) are held more or less motionless within the gesture space.

The current analysis focuses on gestures in the vicinity of the syntactically complete talk spurt end locations; we operationalize this as the 500ms preceding and following the talk spurt boundary, for a total of one second of time surrounding each boundary. The gesture activity was then categorized using the following criteria:

- Is there a gesture occurring before the end of the talk spurt (i.e within the 500ms region preceding the talk spurt boundary)?
- If a gesture occurs in this region, when does it end?
- Is the gesture static (i.e. annotated as a pre- or post-stroke hold gesture phase) or dynamic at the time of the boundary?

2.3. Research questions

We investigate whether gestural prosody is correlated with turn transition, either independently of the acoustic prosody, or in collaboration with it (e.g. in a trading relationship). Since segmental duration and pitch have both been reported to be relevant turn-transition cues, we look for relationships between these acoustic features and speakers' gesture behavior at syntactically complete talk spurt ends.

3. Results

3.1. Spoken turn segments and gesture counts

From the spoken data, we identified 122 syntactically complete talk-spurt ends. Of these, 35 were produced with co-occurring hand gestures in their vicinity. The results reported in sections 3.2 and 3.3 take into account only acoustic data from the turns with co-occurring gesture.

3.2. Gesture offset

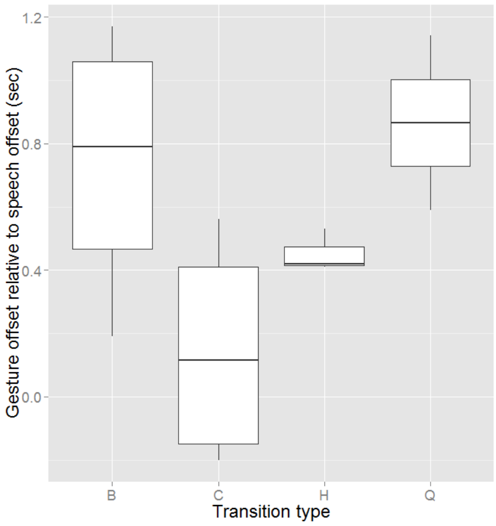

Figure 1: *Offset of gesture compared to offset of speech in different turn transition types. A value of 0 on the y-axis represents gesture offset simultaneous with speech offset. Negative values indicate gesture end preceding speech offset, and positive values gesture end following speech offset. B = Backchannel, C = Change, H = Hold, Q = Question.*

We find an important role of gesture offset timing in relationship to turn transition. Specifically, when speakers gesture in the vicinity of a potential turn boundary location, if that gesture ends before the offset of speech, but within the window space of 500ms preceding the offset, then the speech offset can only be followed by turn change; that is, the following speaker takes up a complete turn (not only a backchannel). This is shown in figure 1.

A visual inspection of the data suggests that end times for gestures at talk spurt ends leading to turn hold are not very variable, coming around 400ms after the offset of speech. An F-test for variance indicates that the variances for the different categories are indeed significantly different (F(117, 121) = 0.048448, p<0.001***).

3.3. Gesture dynamics

While the temporal location of the hand gesture appears to be a valuable cue to turn transition type independently of the acoustic prosodic features, the characteristics of the gesture itself in our data show a relationship with the phonetic characteristics of the co-occurring speech; specifically with segment duration. [16] and [17] report that segment duration at the ends of turn holds is longer than at other potential turn boundary locations. However, in the current data, we find that this segmental length-

Table 1: *Linear mixed model for gesture ends related to time of speech offset. Speaker and segment identity are included as random factors in the model. R^2 = 0.2855088.*

	Estimate	Std. Error	DF	t-value	p-value
(Intercept)	0.082599	0.005538	34.4	14.914	<0.001 ***
transition C	-0.004166	0.003556	834.5	-1.172	0.242 n.s.
transition H	-0.004885	0.003623	800.1	-1.349	0.178 n.s.
transition Q	0.008213	0.005667	794.9	1.449	0.148 n.s.
static	-0.008889	0.009755	627.3	-0.911	0.363 n.s.
transition C*static	0.006996	0.014055	846.7	0.498	0.619 n.s.
transition H*static	0.109828	0.024036	828.4	4.569	<0.001 ***
transition Q*static	0.031269	0.019976	835.0	1.565	0.118 n.s.

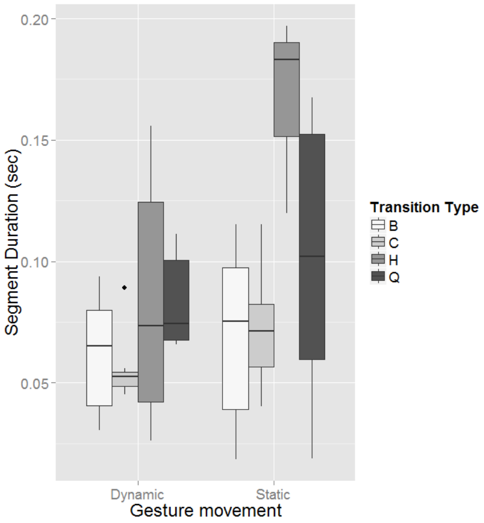

Figure 2: *Segmental duration in speech accompanying gestures. Segmental duration is shown on the left for turn transition types with dynamic accompanying gestures, and on the right for turn transition types with static accompanying gestures. B = Backchannel, C = Change, H = Hold, Q = Question.*

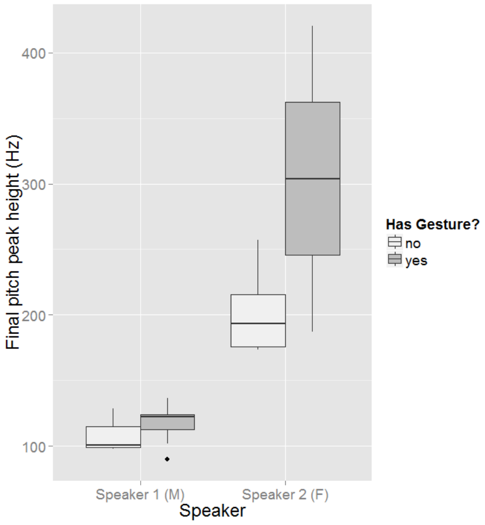

Figure 3: *Final peak pitch height (Hz) for speakers in dialogue 36 when the speaker is or is not also gesturing.*

ening occurs in concert with held gestures; that is, when a co-occurring gesture is static, then it is accompanied by segmental lengthening. However, if a co-occurring gesture is dynamic, then it is not accompanied by segmental lengthening in the current data. This is illustrated in figure 2; statistics are given in table 1.

3.4. Pitch and gesture

In the current data we do not find a relationship between pitch and gesture behavior that differentiates between different turn transition types. However, in one dialogue, there is a relationship between pitch and gesture; specifically, upon visual inspection, both speakers have more variable, and on average higher, pitch in the vicinity of talk spurt ends when they are also gesturing than when they are not. This is the case for both the final pitch peak and the end pitch; since the end pitch data may be distorted due to the presence of glottalization in many tokens, the final pitch peak values for these two speakers are shown in figure 3. T-tests comparing speaker pitch in these contexts did not achieve statistical significance, but this may be because of the small size of the dataset.

4. Discussion

4.1. Gesture timing

Hand gesture timing in the current data was relevant for turn transition, with gestures ending in the 500ms region preceding a syntactically-complete talk spurt end only occurring in turn change locations. This is consistent with previous reports on the timing of gesture at turn ends (e.g. [24]). However, the variation in behavior at locations leading to backchannels, or locations where the talk spurt ends in a question, is consistent with other reports ([23], [25]) that gesture may also be used to invite speech from an interlocutor, and may not be released before such speech has been elicited. Speech in the form of a question naturally invites a response from an interlocutor to fill in a gap, and backchannels can be used to show understanding or involvement on the part of a listener ([30]). Thus, these contexts may differ in structure from the simple change cases, where the preceding turn is syntactically complete but does not necessarily call for a specific response. The end of the gesture before the end of speech in turn change contexts may indicate a more open scenario for the continuation of speech by the other dialogue partner, whereas backchannel and question contexts are more constrained being marked by gesture activity continuing across the talk-spurt boundary.

The finding that gestures for turn hold end with very little variation around 400ms after the offset of speech suggests a duration for the transition space. Specifically, this may be the length of time necessary for the speaker to retain his/her turn in an otherwise “neutral” setting by gesturally preventing the other speaker from beginning. The greater variation in gesture duration in other contexts may be due to the invitation of speech or the lack of completion of a conversational activity, as described above. Of course, it may also simply result from a choice to begin talking regardless of an interlocutor's expressed desire about the continuation of the conversation. Regardless, it is interesting to see the close construction of turn hold contexts in terms of the timing of gesture ends in the transition space, as well as the phonetic structure in relation to the gesture form, as discussed in the next section.

4.2. Holding gestures and lengthened speech

We find that Swedish speech in hold contexts with accompanying gesture is only lengthened when the gesture is itself static; that is, held speech accompanies held gesture. This corresponds with [37]'s analysis of American Sign Language gestures, in which slowing and holding of gestures are associated with phrase-finality but are not treated as simple pauses. On the other hand, segment duration in turn holds associated with dynamic gesture is extremely variable, suggesting that segmental lengthening in this context is also possible, but not necessary. One possibility requiring further examination is that the features of the gestural movement (location, direction, or velocity) may be related to segmental duration in a more specific way; a future analysis will use the motion capture data in Spontal to address this question.

It is also possible that there are different kinds of turn holds, and that our four-way distinction of syntactically complete talk-spurt boundaries is not a sufficient analysis. Segmental lengthening (and gesture) could be relevant at some kinds of turn holds and not others. [23], [24] and [25] report that holding gestures may occur even across speaker transitions as a way of indicating that some conversational action is not yet complete. Thus it may be necessary to distinguish not only syntactically complete talk-spurt ends but also "actionally complete" ones in order to gain an accurate picture of the role gesture and acoustic prosody play in facilitating turn holds.

4.3. Pitch

In the current data, pitch did not show a clear relationship with turn-taking behavior. However, there was some indication that pitch and gesture still have a relationship; specifically, the participants in one dialogue appeared to have higher pitch when their speech was accompanied by gesture. This dialogue is impressionistically the one in which the participants are most engaged. A possible interpretation for the more variable/higher pitch values accompanying gesture is that these locations are specifically those at which the speakers are particularly emotionally engaged. This may reflect a greater degree of bodily activation in general. However, this explanation has not been investigated in further detail in the present study.

The fact that such behavior was not found for participants in other dialogues should not be interpreted as contradicting this finding, since those participants gestured much less frequently, and one produced such an extreme amount of glottalization that almost no pitch values were available for his speech. Thus the question of the relationship of pitch to gesture at Swedish talk-spurt ends remains an open one.

4.4. Combined system

We have found evidence for co-occurring prosodic and gestural cues to turn transition, as well as for independent behavior of gesture. This suggests that, just as multiple acoustic features can contribute to prosodic precepts (e.g. tonal activity and lengthening at phrase boundaries), we must also take gesture into consideration as part of the prosodic system. This idea has been proposed by e.g. [27] and is consistent with other observations on the relationship between gesture and prosody, e.g. [38]. The current data give us a first approximation of how we might consider gestures as part of a prosodic system of turn-taking, and not simply as an independently-functioning cue, but further research is necessary to better understand the systematic relationships present.

5. Conclusions

We find that variations in hand gesture timing are related to turn transition structure in Swedish conversation, and that a complex relationship appears to exist between gesture type and increased final lengthening in turn hold contexts. A possible relationship of pitch variation to gesture presence is also reported. On the basis of these data, we suggest that prosodic cues to turn-taking, and likely to other conversational functions, must be investigated in parallel with gesture cues, and with an eye to the likelihood of their forming a single system, rather than two separate ones.

6. Acknowledgements

The work reported here has been funded by the Bank of Sweden Tercentenary Foundation (P12-0634:1) and the Swedish Research Council (VR 2010-4646).

7. References

[1] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organisation of turn-taking for conversation," *Language*, vol. 50, pp. 696–735, 1974.

[2] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon, and S. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 26, pp. 10 587–10 592, 2009.

[3] D. Schaffer, "The role of intonation as a cue to turn taking in conversation," *Journal of Phonetics*, vol. 11, pp. 243–57, 1983.

[4] P. Auer, "On the prosody and syntax of turn-continuations," in *Prosody in conversation: interactional studies*, E. Couper-Kuhlen and M. Selting, Eds. Cambridge, UK: Cambridge University Press, 1996, pp. 57–100.

[5] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speakers turn: a cognitive cornerstone of conversation," *Language*, vol. 82, no. 3, pp. 515–535, 2006.

[6] J. Local, J. Kelly, and W. Wells, "Towards a phonology for conversation: turn-taking in Tyneside English," *Journal of Linguistics*, vol. 22, pp. 411–437, 1986.

[7] M. Selting, "On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation," *Pragmatics*, vol. 6, pp. 357–388, 1996.

[8] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics*, vol. 31, pp. 251–276, 2003.

[9] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetics Association*, vol. 31, pp. 139–152, 2001.

[10] J. Kane, I. Yanushevskaya, C. de Looze, B. Vaughan, N. Chasaide, and A., "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *Proceedings of 15th Interspeech, Singapore*, 2014, pp. 333–337.

[11] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs," *Language and Speech*, vol. 41, pp. 295–321, 1998.

[12] A. Gravano and J. Hirschberg, "Turn-yielding cues in task-oriented dialogue," in *Proceedings of SIGDIAL 2009, Queen Mary University of London, UK*, 2009, pp. 253–261.

[13] ——, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, pp. 601–634, 2011.

[14] A. Hjalmarsson, "The additive effect of turn-taking cues in human and synthetic voice," *Speech Communication*, vol. 53, pp. 23–35, 2011.

[15] M. Zellers, "Duration and pitch in perception of turn transition by Swedish and English listeners," in *Proceedings of FONETIK 2014, Stockholm, Sweden*, M. Heldner, Ed., Jun. 2014.

[16] ——, "Pitch and lengthening as cues to turn transition in Swedish," in *Proceedings of 14th Interspeech, Lyon, France*, 2013, pp. 248–252.

[17] ——, "Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish," submitted.

[18] A. Hjalmarsson and K. Laskowski, "Measuring final lengthening for speaker-change prediction," in *Proceedings of 12th Interspeech, Florence, Italy*, 2011.

[19] M. Heldner and M. Włodarczak, "Pitch slope and end point as turn-taking cues in Swedish," in *Proceedings of ICPhS 2015, Glasgow, Scotland*, Aug. 2015, pp. 10–15.

[20] J. Edlund and J. Beskow, "Pushy versus meek - using avatars to influence turn-taking behaviour," in *Proceedings of Interspeech 2007, Antwerp, Belgium*, 2007.

[21] ——, "Mushypeek – a framework for online investigation of audiovisual dialogue phenomena," *Language and Speech*, vol. 52, pp. 351–367, 2009.

[22] C. Heath, "Talk and recipiency: sequential organization in speech and body movement," in *Structures of social action: studies in conversation analysis*, J. M. Atkinson and J. Heritage, Eds. Cambridge, UK: Cambridge University Press, 1984, pp. 247–265.

[23] J. Streeck and U. Hartge, "Previews: Gestures at the transition place," in *The Contextualization of Language*, P. Auer and A. di Luzio, Eds. Amsterdam: Benjamins B.V., 1992, pp. 135–158.

[24] L. Mondada, "Multimodal resources for turn-taking: pointing and the emergence of possible next speakers," *Discourse Studies*, vol. 9, no. 2, pp. 194–225, 2007.

[25] R. O. Sikveland and R. Ogden, "Holding gestures across turns: moments to generate shared understanding," *Gesture*, vol. 12, no. 2, 2012.

[26] F. Quek, D. McNeill, R. Bryll, S. Duncan, X. F. Ma, C. Kirbas, K. E. McCullogh, and R. Ansari, "Multimodal human discourse: gesture and speech," *ACM Transactions on Computer-Human Interaction*, vol. 9, no. 3, 2002.

[27] D. Gibbon, "Gesture theory is linguistics: modelling multimodality as prosody," in *Proceedings of PACLIC 23 Conference, Hong Kong*, 2009.

[28] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proceedings of LREC 2010, Valetta, Malta*.

[29] K. Laskowski, *Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation*. Doctoral dissertation, Carnegie Mellon University, 2011.

[30] R. Gardner, *When listeners talk: response tokens and listener stance*. Amsterdam: John Benjamins, 2001.

[31] D. House, "Final rises and Swedish question intonation," in *Proceedings of Fonetik 2004, Stockholm University, Sweden*, 2004.

[32] ——, "Phrase-final rises as a prosodic feature in wh-questions in Swedish humanmachine dialogue," *Speech Communication*, vol. 46, pp. 268–283, 2005.

[33] Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, "ELAN," http://tla.mpi.nl/tools/tla-tools/elan/, 2015.

[34] H. Brugman and A. Russel, "Annotating Multimedia/Multi-modal resources with ELAN," in *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*.

[35] D. McNeill, *Hand and mind: What gestures reveal about thought*. Chicago, IL: The University of Chicago Press, 1992.

[36] S. Kita, I. van Gijn, and H. van der Hulst, "Gesture and sign language in human-computer interaction," in *Lecture Notes in Computer Science*, I. Wachsmuth and M. Frölich, Eds., 1998, vol. 1371, pp. 23–35.

[37] M. E. Tyrone, H. Nam, E. Saltzman, G. Mathur, and L. Goldstein, "Prosody and movement in American Sign Language: a task-dynamics approach," in *Proceedings of Speech Prosody 2010, Chicago, IL, USA*, 2010.

[38] B. Granström and D. House, "Audiovisual representation of prosody in expressive speech communication," *Speech Communication*, vol. 46, pp. 473–484, 2005.