Statistical Unpredictability of F0 Trajectories as a Cue to Sentence Stress

Sofoklis Kakouros (sofoklis.kakouros@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Okko Räsänen (okko.rasanen@aalto.fi)
Department of Signal Processing and Acoustics, Aalto University,
PO Box 13000, AALTO, Finland

Abstract

This paper introduces a hypothesis that the perceived sentence stress in speech is related to the unpredictability of prosodic features, thereby capturing attention of the listener. In order to study this idea, a computational model was designed that learns the statistical structure of temporal F0 trajectories from continuous speech data without supervision using n-gram statistics. When the model output is compared to human perception of stress on a set of novel utterances, the low-probability points of the F0 trajectories show high correlation with the moments of subjective perception of stress. The result gives support to the idea that perceptual attention and unpredictability of sensory stimulus are mutually connected, and suggests that stress perception can be learned with similar statistical learning mechanisms that are considered to play a central role in early word segmentation.

Keywords: sentence stress; prosody; pitch; attention; statistical learning

Introduction

Sentence stress is a universal property of speech where a word or multiple words of an utterance are given a special emphasis in the message in order to facilitate the listener's perception and draw attention to these aspects of the content (e.g., Cutler & Foss, 1977). It is widely accepted that stress is correlated with prosodic features such as pitch, loudness, and timing (Imoto et al., 2002; Cutler & Foss, 1977; see also Cutler, Dahan & Donselaar, 1997, for a review). Also, the acoustic correlates of stress seem to be relatively universal across languages although the specific realizations of stress patterns may vary substantially from one language to another with respect to the underlying linguistic content (Endress & Hauser, 2010). The characteristics of prosody and prominence are also similar in infant-directed speech (IDS) and adult-directed speech (ADS), although the prosodic modulation is typically exaggerated in the IDS and the stressed words in IDS may exhibit more systematic relative positioning in the utterance in comparison to the ADS (Endress & Hauser, 2010; Ferguson, 1964; Fernald & Mazzie, 1991; Grieser & Kuhl, 1988; Remick, 1976).

However, despite the well documented findings on acoustic and linguistic characteristics of stress, it is not well understood why listeners pay attention to these specific aspects of the acoustic signal and, on the other hand, how the variability in the stress patterns across different languages are related to the perceptual processing. This also poses the question of learnability versus innateness of stress perception: If stress perception is innate, it is likely that there are universal physical characteristics that define stressed words with respect to unstressed ones. On the other hand, if stress perception is learned, the relevant question is then what is actually learned from the signals, how it is learned, and how does this learning of prosody relate to the other aspects of speech perception. The problem regarding the nativist (or "hard-wired") approach is that it has difficulties in explaining the differences in stress use across languages, talkers, and speaking styles. On the other hand, perceptual learning of stress cannot be based on explicit instruction since, at least according to the knowledge of the authors, language learners rarely receive feedback for their supra-segmental perceptual processing of language.

The role of learning in stress perception is especially relevant in the context of language acquisition research. For example, word-level stress patterns are known to be relevant for early word segmentation (Endress & Hauser, 2010; Jusczyk, Cutler & Redanz, 1993; Peters, 1983; Thiessen, Hill & Saffran, 2005). Similarly, it can be hypothesized that one role of sentence stress in infant-directed speech (IDS) is to provide attentional cues to the child regarding the important content words in the message (Fernald & Mazzie, 1991). For example, focus on specific words may facilitate cross-situational learning between acoustic word forms and their referents by constraining the number of relevant candidate words in the present utterance.

As for the computational modeling of stress detection and perception, previous approaches have primarily focused on supervised learning of the relationship between prosodic (acoustic) features and the stressed and/or non-stressed units of speech (Chaolei, Jia & Shanhong, 2007; Imoto, Dantsuji & Kawahara, 2000; Imoto et al., 2002; Lai et al., 2006; Minematsu et al., 2002; Ringerval et al., 2011; Rosenberg & Hirchberg, 2009). For example, Imoto et al. (2002) proposed a two-stage model where a weighted combination of F0, signal power, and Mel-frequency cepstral coefficients (MFCCs) are used in a hidden-Markov model (HMM) to determine the presence or absence of stress in a word, followed by a more close evaluation of the stress level. Another approach was presented by Minematsu et al. (2002) who proposed a technique for modeling stressed and unstressed syllables in speech by analyzing the relative differences of acoustic features between consecutive syllables. These differences were then used to determine whether a specific syllable is characterized as stressed or unstressed. They also used supervised HMM training to learn the associations between the acoustic features and human-made stress annotation. Finally, an approach combining the acoustical features with the linguistic properties of the utterances was proposed by Lai et al. (2006). Their method utilizes three layers of classifiers where the first two layers assign stress labels to content words and unstressed labels to function words while the third classifier performs stress labeling based on the acoustic features of the utterance. In all of the above studies, the stress is detected in terms of presence of specific acoustic features, their differences between subsequent linguistic units, or in terms of linguistic content of the signal, and the connection between these cues and presence of stress is learned in a supervised manner using machine learning algorithms.

In contrast to the supervised approaches, sentence level stress can also be examined from purely unsupervised point of view. More specifically, we put forward a hypothesis that the *statistical unpredictability of the prosodic features is the main carrier of stress in speech*. Therefore no supervised associative learning is required. The idea is inspired by the cognitive foundations of attention where the primary role of the attentional system is to select novel or otherwise significant information from the environment (Broadbent, 1958; Treisman, 1964). In other words, attention can be seen as a mechanism for allocating active sensory- and learning resources for input that is not anticipated by our existing predictions regarding the state of the surrounding world. It also seems that the the human brain is wired to react to expectations and their violations in the sensory input (see Itti & Baldi, 2009, and references therein).

In the given framework, potential points of stress in speech can be hypothesized to be the points where the prosodic features deviate from their expected outcomes in the given context. These deviations can be actively controlled by the talker who is relatively free to choose the suprasegmental acoustic parameters as a function of position in the utterance while the listener can only rely on the previously learned a priori expectations for these parameters. Importantly, the listener's expectations can be learned from exposure to speech without any supervision, i.e., without access to a ground truth on the degree of stress in the heard words. Instead, a statistical model of the typical behavior of the prosodic features is sufficient, and it is now widely accepted that human infants and adult are sensitive to regularities in the sensory input (e.g., Romberg & Saffran, 2010, and references therein).

In the current work, we study the statistical learning hypothesis by modeling the fundamental frequency contours of continuous speech and comparing the model output to human perception of sentence stress in the same utterances. More specifically, we test whether the regions of low-probability in the F0 contours match with the human stress, showing that there is substantial correlation between the two.

Material

The CAREGIVER Y2 UK corpus (Altosaar et al., 2010) was used in the study. The style of speech in CAREGIVER is enacted IDS spoken in continuous UK English, simulating a situation where a caregiver is talking to a child regarding a number of important objects and events in a shared interaction scene, but recorded in high-quality within a noise-free anechoic room. In addition to a set of 50 unique keywords, the speech material contains a number of verbs and function words used in the surrounding carrier sentences, yielding a total vocabulary of 80 words. The vocabulary is statistically balanced over the keywords so that the predictive relationships between keywords and keyword pairs (e.g., an adjective and a noun) are minimized. The talkers were not separately instructed on the use of prosody or stress beyond that they were asked to read the text prompts, paired with visual stimuli, as they would talk to their own children (see Altosaar et al., 2010, for details).

In overall, CAREGIVER UK Y2 contains 2397 sentences from each main talker. A subset of 300 unique utterances were chosen for the listening tests from one male and female talker (*Speakers 3 and 4*), yielding a total of 600 utterances. All single-word utterances were excluded from the data, leading to an average of 5.9 words per sentence. This set of utterances is referred to as the *test set*, as it was also used to probe the performance of the studied statistical F0 model (see the Methods section).

As for the training of the statistical model, 2000 sentences per talker were used (i.e., 4000 in total). None of these *training set* sentences were present in the above test set.

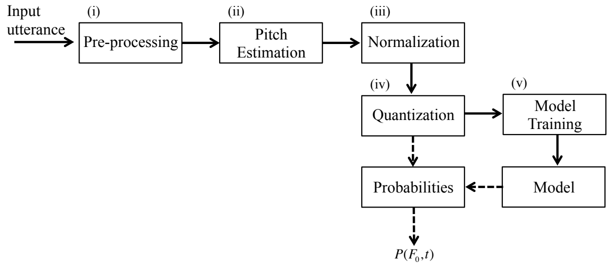

Figure 1: Overview of the proposed model.

Methods

The study of the statistical learning hypothesis in stress perception requires: 1) the collection of a reference annotation representing human perception of stress in the test set, and 2) a computational model that can learn the statistical regularities of the F0 trajectories from the training data and then evaluate the points of unpredictability on the same data that human listeners were exposed to.

Stress Annotation

In order to create a reference annotation of sentence stress against which the model could be tested, an annotation tool with a graphical user interface (GUI) was created for MATLAB. The GUI plays each utterance through headphones, displays the list of spoken words in a temporally ordered list on a computer screen, and then prompts the user to choose the words that were perceived as stressed using a computer mouse as the controller. For each utterance, the test subject can select zero or more words as stressed. The test subject can also listen to each sentence as many times as he/she wishes.

The listening tests were performed in a sound-isolated listening booth using Sennheiser HD650 headphones fed through Motu Ultralink MK3 audio interface. The listeners were able to take a break any time between the utterances.

Annotation data for the current study were collected from a total of thirteen test subjects (7 male, 6 female) from the age range of 20-30 years. Nine of the participants were L1 Finnish speakers, one British English, one Greek, one Russian and one Spanish. English and Swedish represented the majority of the L2 and L3 languages among the listeners. Despite the various L1 and L2 combinations, each listener was considered as experienced English user. On average, the task took approximately 1.5 hours per listener.

Statistical Model of Prosodic Trajectories

The overall aim was to build an unsupervised statistical model of the temporal evolution of the F0 trajectories based on the training set of utterances and then to detect word stress in terms of the unpredictability of the F0 during the test set. The overall model consists of the following steps: (i) signal pre-processing, (ii) F0 estimation, (iii) F0 normalization, (iv) F0 quantization, and (v) n-gram parameter estimation (during training) or n-gram probability computation (during testing; see Figure 1).

In the pre-processing step, the speech data were downsampled from 44.1 kHz to 8 kHz. Then the F0 estimation and voicing detection was performed for each utterance using the YAAPT-algorithm (Zahorian & Hu, 2008). "Filler" F0 contours for unvoiced segments were generated by linear interpolation from the surrounding voiced F0 values. This was done in order to ensure temporal continuity of the data without introducing any new information to the signal that could cue stress or absence of stress in the signals (see Fig. 2, second panel).

In order to ensure F0 comparability across different utterances and the two talkers, the original absolute frequency F0 contours were normalized according to

$$F0_{\mathrm{N}}(t) = \frac{F0(t) - \min(F0)}{\max(F0) - \min(F0)} \qquad (1)$$

where min(F0) and max(F0) refer to the minimum and maximum of the F0 during the given utterance, effectively scaling the $F0_N$ between 0 and 1 (see Imoto et al., 2002).

Finally, in order to enable probabilistic modeling of the normalized F0 contours, the $F0_N$ were quantized into 32 discrete amplitude levels that were estimated using the k-means algorithm with a random sampling initialization. The number of levels was selected as a compromise between the best possible approximation of the F0 contours without ending up with too sparse statistics for the different levels.

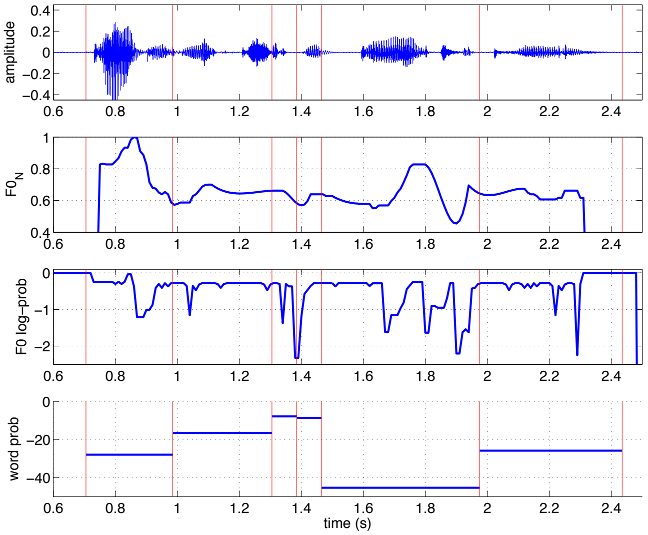

Figure 2: Example output of the algorithm for the utterance "*Daddy looks at the* ***dirty*** *car*", with the word "*dirty*" annotated as stressed by the majority of the annotators. Top panel: The original signal waveform. Second panel: The normalized $F0_N$ contour. Third panel: The corresponding 4-gram log-probabilities (with 3-point median filtering for improved visual clarity). Bottom panel: Cumulative word log-probabilities across the entire word duration. Vertical lines denote word boundaries.

As for the statistical modeling of the temporal evolution of the discretized F0 values, standard n-grams were chosen for the purpose due to their computational simplicity and ease of interpretation. The analysis was limited to n-gram orders of n = 2, 3 and 4, where bi-grams (n = 2) correspond to the shortest temporally ordered segments available while the four-grams (n = 4) are the longest recurring sequences for which probabilities can be reliably estimated from the data.

The probabilities for the n-grams were computed using the maximum-likelihood estimator, i.e., using the relative frequencies of the n-tuples occurring in the training set.

$$P(F0_i \mid F0_{i-1},\cdots,F0_{i-N+1}) = \frac{C(F0_i,F0_{i-1},\cdots,F0_{i-N+1})}{C(F0_{i-1},\cdots,F0_{i-N+1})} \qquad (2)$$

$$P'(\mathrm{F0},t) = \log(P(F0_t \mid F0_{t-1},\cdots,F0_{t-N+1})) \qquad (3)$$

In the equations, C denotes the frequency counts of the discrete F0 n-tuples in the training data and F0 refers to the discretized F0 values in the range 1–32.

During the test phase, steps (i)-(iv) were the same as in the training. The probability curve of the F0 as a function of time (Figure 2, third panel) was computed for each utterance using the Eq. (3). Then the logarithm of the probabilities was taken to avoid numerical instability. Also, in order to avoid log(0) for previously unseen F0 contours, zero probability contours were floored to P' = log(0.00001).

In order to simulate the listening test task of choosing N stressed words out of the M possibilities in each sentence, the instantaneous F0 probabilities were converted into word-specific probability scores $S(w)$ by simply summing the log-probabilities of the F0 trajectory over the duration of each word (Figure 2, bottom).

$$S(w) = \sum_{t=t_1}^{t_2} P'(\mathrm{F0}, t) \qquad (4)$$

where t_1 and t_2 are the known word boundaries in time.

Finally, the stress hypotheses for words were generated from the word scores by choosing the words w_y that had their overall score $S(w_y)$ below a threshold r_i. The r_i was defined dynamically according to

$$r_i = \mu_i - \sigma_i\lambda \qquad (5)$$

i.e., as λ standard deviations σ_i from the mean μ_i of the scores across all words in the same utterance *i*. Value of the free parameter λ was varied in the experiments in order to observe the behavior of the model as a function of the detection threshold. It should be noted that the use of a global fixed threshold across all utterances was also studied and it was found to lead to similar performance with Eq. (5).

Evaluation

In order to measure inter-annotator agreement rate in the listening test and in order to compare model output to the human annotations, the standard Fleiss kappa (Fleiss, 1971) measure was used. In essence, the Fleiss kappa measures the degree of agreement between two or more annotators on a nominal scale $\kappa \in [-1,1]$. It takes into account the underlying distribution of the ratings, yielding $\kappa = 0$ if the number of agreements is equal to what is expected based on chance-level co-occurrences in the data. In the current study, the Fleiss kappa was measured on a word-level, i.e., each word occurring in the test set was considered as a binary decision between non-stressed and stressed. The overall agreement rate across all words in the test set were used as the primary evaluation measure. As for the listening test data, we measured the overall kappa value across all thirteen annotators and the pair-wise kappas for each possible pair of annotators. In order to evaluate the model, the stress hypotheses of the model were compared in a pair-wise manner with the markings of all individual annotators and the average across all model-annotator-pairs was computed.

In order to understand chance-level performance in the task, two different random baseline results were also generated. In the *basic baseline*, each word was randomly assigned as either stressed or unstressed with the constraint that each utterance receives the same number of stress words as hypothesized by the model with a given detection threshold λ. In the so-called *duration baseline*, the process was otherwise similar but, instead of sampling from a uniform distribution, the probability of a word being assigned as stressed was linearly proportional to the duration of the word. The duration baseline represents the performance achieved by simply integrating random signal (random probabilities) across the word lengths with longer words thereby leading to lower overall probabilities (cf., Eq. (4)). Duration baseline therefore also provides indirect evidence of the role of duration in stress perception, as the duration cannot be directly represented as a signal feature in the current type of temporal model. Both baselines were computed across 50 iterations of random sampling.

Results

Annotation Analysis

The set of 600 test signals were stress annotated by thirteen separate annotators. The overall Fleiss kappa across all annotators was κ = 0.4, which translates into mean agreement rate of 85.7% for individual word tokens. On average, a total of 23.6% (±5.6%) of all words were considered as stressed. As for the pair-wise agreements between annotators, the average agreement was κ = 0.39 with a standard deviation of 0.12, a minimum of κ = 0.12, and a maximum of κ = 0.65. This indicates that there is notable variation across the annotators, some of the listeners sharing a very similar perception of stress across a large number of utterances while some others have very different view on what is stressed or not. The average agreement of 0.4 is significantly above chance level and is at the boundary of “fair” and “moderate” agreement on the Landis & Koch (1977) scale. It is also the same agreement rate observed in two other studies of prominence perception using American English (Mo, Cole & Lee, 2008; You, 2012). In overall, the results from the listening tests confirm that the sentence-level prominence is not a clear unanimous phenomenon, but more like a fuzzy continuum from unstressed to stressed words. Still, there is a systematic tendency among listeners to perceive specific parts of the signals as stressed (more detailed analysis of the listening test data is beyond the scope of the current paper).

Model Simulations

The statistical model of F0 was evaluated for a number of detection thresholds by varying the parameter λ of Eq. (5). The overall results with the model can be seen in Figure 3 where the result is averaged across the three studied n-gram orders (*n* = 2, 3, 4). All results are pooled across the male and female talker of the test set. The best correspondence to manually annotated stress words is obtained for a threshold of 1/2 standard deviations below average F0 probability during the word, leading to mean pair-wise agreement of κ = 0.39 with the annotators. For the individual n-gram orders, the agreements are κ = 0.41, 0.40, and 0.38 for the bi-, tri-, and four-grams, respectively.

As can be observed from Figure 3, the average statistical model agreement with the behavioral data is basically equal to the inter-rater agreement level and significantly above both chance-level performances. As expected, the uniformly sampled random baseline achieves κ = 0.00. In contrast, the duration baseline reaches a slight agreement of κ = 0.16.

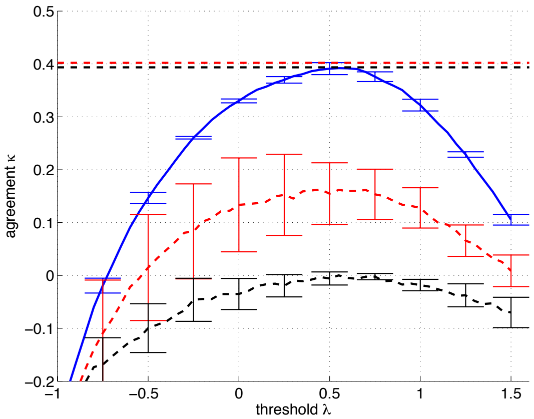

Figure 3: Pair-wise Fleiss kappa between the model output and the human listeners as a function of detection threshold λ. The blue solid line shows the mean and standard deviation of model performance across n-grams orders n = 2, 3, and 4. The black dashed line shows the basic chance level performance and the red dashed line at the middle shows the chance-level performance when the word durations are taken into account. The red and black horizontal dashed lines indicate the overall and mean pair-wise kappas across the annotators, respectively.

Figure 4 shows the pair-wise agreement of the model output with each annotator together with the corresponding duration baselines. As can be observed, the notable variation between annotators is also evident in the model output comparisons, the model agreeing with annotator 11 with above κ = 0.6 using tri-grams while agreement with annotator 8 is only slightly above κ = 0.2.

In general, the current results provide strong support to the statistical learning hypothesis as the agreement rate of the algorithm with human listeners is comparable to that of any human listener.

Conclusions

In this work, we formulated a hypothesis that the perception of sentence stress in speech is related to the unpredictability of the acoustic correlates of prosody. This is in contrast to the classical approach where stress is defined in relation to specific configurations of acoustic feature values and assuming that the listener knows this relationship in advance. The unpredictability hypothesis was tested by modeling the temporal evolution of fundamental frequency with a simple unsupervised statistical model. The model marks words as prominent if the F0 trajectory during them is unlikely given the earlier learned expectations. As a result, the model shows high agreement with human perception on the same task and thereby provides the first evidence to the idea that stress perception can be learned from statistical regularities of speech.

The idea of learning supra-segmental linguistic cues from statistical regularities has interesting parallels to the ongoing research in early language acquisition. For example, infant's capability for word segmentation has been discussed in the context of statistical learning of transitional probabilities (TPs) between linguistic units such as phones or syllables (see Romberg & Saffran, 2010, for a review). However, it is known that prosody also helps in word segmentation (e.g., Johnson & Jusczyk, 2001), with the statistical learning and prosody perception being treated as two distinct mechanisms. However, the current work suggests that a statistical learning mechanism could account for the sensitivity to both types of cues with the only difference in the acoustic features that are being learned. In the TP-based word segmentation, the focus is on the statistical structure of linguistically relevant features such as formant frequencies or the overall spectrum of speech (see Räsänen, 2011, 2012). In contrast, the perception of prosody may be driven by the statistical structure of the suprasegmental features with the degree of predictability in the prosody modulating the attentional focus of the listener.

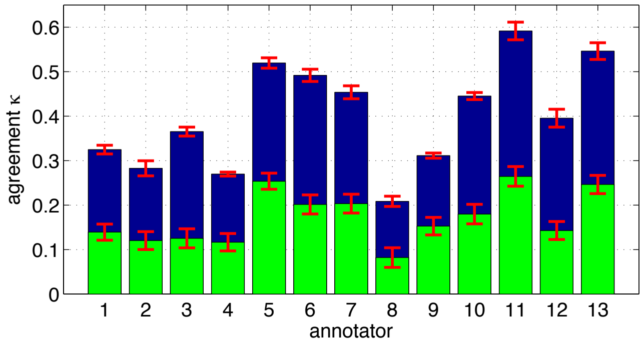

Figure 4: Pair-wise agreement rates between the model output and each individual annotator. The result is the mean and SD of the three evaluated n-gram orders. Green bars at the bottom show the duration-baseline agreement levels.

However, more work is needed to consolidate the current findings. First of all, the experiment should be extended to adult-directed speech in order to see whether the findings persist, although the acoustic features of prominence in ADS and IDS should be similar but simply of different magnitude (e.g., Song, Demuth & Morgan, 2010). Also, it is currently unknown how much exposure is needed to form the expectations for the prosodic trajectories (e.g., short-term vs. long-term memory), and whether these expectations generalize across talkers and across communicative contexts. In addition, we have so far considered only one of the acoustic correlates of prosodic features, namely F0. Other features such as loudness (energy), spectral tilt, or the full wide-band spectrum of the signal should be investigated using the same approach. These studies are beyond the scope of the current paper but will be addressed in the future work.

Acknowledgements

This research was funded by the Data to Intelligence (D2I) project of Tekes and by the Academy of Finland.

References

Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H. (2010). A Speech Corpus for Modeling Language Acquisition: CAREGIVER. *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), Malta, pp. 1062–1068.

Broadbent, D. E. (1958). *Perception and Communication*. New York: Pergamon.

Chaolei, L., Jia, L., & Shanhong, X. (2007). English Sentence Stress Detection System Based on HMM framework. *Applied Mathematics and Computation*, 185, 759–768.

Cutler, A., & Foss, D. J., (1977). On the Role of Sentence Stress in Sentence Processing. *Language and Speech*, 20, 1–10.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, 40, 141–201.

Endress, A. D., & Hauser, M. D. (2010). Word Segmentation with Universal Prosodic Cues. *Cognitive Psychology*, 61, 177–199.

Ferguson, C. A. (1964). Baby talk in six languages. *American Anthropologist*, 66, 103–114.

Fernald, A., & Mazzie, C. (1991). Prosody and Focus in Speech to Infants and Adults. *Developmental Psychology*, 27, 209–221.

Fleiss, J. L. (1971). Measuring norminal scale agreement among many raters. *Psychological Bulletin,* 76, 378–382.

Grieser, D. L., & Kuhl, P. K., (1988). Maternal Speech to Infants in a Tonal Language: Support for Unviersal Prosodic Features in Motherese. *Developmental Psychology*, 24, 14–20.

Imoto, K., Dantsuji, M., & Kawahara, T. (2000). Modelling of the Perception of English Sentence Stress for Computer-Assisted Language Learning. *Proceedings of Interspeech*, Beijing, China, pp. 175–178.

Imoto, K., Tsubota, Y., Raux, A., Kawahara, T., and Dantsuji, M. (2002). Modeling and Automatic Detection of English Sentence Stress for Computer-Assisted English Prosody Learning System. *Seventh International Conference on Spoken Language Processing*, pp. 749–752.

Itti, L., & Baldi, P. (2009). Bayesian Surprise Attracts Human Attention. *Vision Research*, 49, 1295–1306.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44, 548–567.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Infants' Preference for the Predominant Stress Patterns of English Words. *Child Development*, 64, 675–687.

Lai, M., Chen, Y., Chu, M., Zhao, Y., & Hu, F. (2006). A Hierarchical Approach to Automatic Stress Detection in English Sentences. *International Conference on Acoustics, Speech and Signal Processing (ICASSP'2006)*, Toulouse, France, pp. 753–756.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement of categorical data. *Biometrics*, 33, 159–174.

Minematsu, N., Kobashikawa, S., Hirose, K., & Erickson, D. (2002). Acoustic Modeling of Sentence Stress Using Differential Features Between Syllables for English Rhythm Learning System Development. *Seventh International Conference on Spoken Language Processing (ICSLP'2002)*, pp. 745–748.

Mo, Y., Cole, J., & Lee, E-K. (2008). Naïve listeners' prominence and boundary perception. *Proc. 4th Speech Prosody*, Campinas, Brazil, pp. 735–738.

Peters, A. M. (1983). *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Remick, H., (1976). Maternal speech to children during language acquisition. In W. von Raffler-Engel & Y. Lebrun (Eds.), *Baby talk and infant speech* (pp. 223–233). Amsterdam: Swets & Zeitlinger.

Ringeval, F., Demouy, J., Szaszák, G., Chetouani, M., Robel, L., Xavier, J., Cohen, D., & Plaza, M. (2011). Automatic Intonation Recognition for the Prosodic Assessment of Language-Impaired Children. *IEEE Transactions on Audio, Speech, and Language Processing*, 19, 1328–1342.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Review of Cognitive Science,* 1, 906–914.

Rosenberg, A., & Hirschberg J. (2009). Detecting Pitch Accents at the Word, Syllable and Vowel Level. *Human Language Technology Conference of the North American Chapter of the ACL*, Boulder, Colorado, pp. 81–84.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, 120, 149–176.

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions. *Speech Communication*, 54, 975–997.

Song, J.Y., Demuth, K., & Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *Journal of the Acoustical Society of America*, 128, 389–400.

Thiessen E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-Directed Speech Facilitates Word Segmentation. *Infancy*, 7, 53–71.

Treisman, A. M. (1964). The effect of irrelevant material on the efficiency of selective listening. *The American Journal of Psychology*, 77, 533–546.

Zahorian, S. A., & Hu, H. (2008). A spectral/temporal method for robust fundamental frequency tracking. *Journal of the Acoustical Society of America*, 123, 4559–4571.

You, H-J. (2012). Determining prominence and prosodic boundaries in Korean by non-expert rapid prosody transcription. *Proc. 6th Speech Prosody*, Shanghai, China.