# COMPARISON OF PROSODIC FEATURES IN SWEDISH AND FINNISH IDS/ADS SPEECH

Okko Räsänen
Toomas Altosaar
Unto K. Laine

***Abstract***

A set of speech corpora, spoken in both infant-directed speech (IDS) and adult-directed speech (ADS) modes containing similar sentences in terms of complexity, was collected in Sweden and Finland in 2007. Prompts for each language were designed by selecting 10 object words and 10 carrier sentences for each language to produce approximately 100 different sentences. Each sentence was spoken 10 times in ADS mode and 10 times in IDS mode, for a total of 2000 sentences per speaker. Four speakers, typically two pairs of married couples from each language, were used to produce a total of 8000 utterances per language. The corpora, that also included Dutch and English, were collected to aid in the exploration of new architectures for speech recognition, as is currently taking place in the ACORNS[1] project

This study describes the main prosodic differences found in IDS/ADS speech from the collected Swedish and Finnish material. Statistical analysis of the following main features are reported:

- F0 histograms: these may reveal speaking style differences in the F0-contours.
- Spectral tilting and its fluctuation: this also is affected by the voice source and the "softer" speaking style encountered with IDS.
- Segmental durations and their variations: average speaking rate and indications of monotonic/non-monotonic speaking style.
- Differences in spectral properties: the signals are analyzed in frames of 10 ms and the MFCC-coefficient vectors clustered in a vector space. The cluster centres reveal average spectral properties and differences in speaking style that can be detected as differences in the number of clusters found as well as differences of their mean vectors.

[1] ACORNS (Acquisition of Communication and Recognition Skills), an EU-funded Sixth Framework Programme, Priority [2], Future and Emerging Technologies, project, 2007–2009.

The analysis is performed for four Swedish and four Finnish speakers (two males and two females for each language). Tests are carried out to detect any common characteristics between IDS and ADS speech for these two languages.

## *1 Introduction*

It is well confirmed that so-called *parentese*, or infant-directed speech (IDS), differs from casual adult-to-adult conversational language (adult-directed speech, ADS). This dissimilar use of language depending on the audience is most often an unconscious action, but infant-directed speech seems also to be a universal phenomenon. Infant-directed speech is often described as more vivid, lexically limited, and more clearly articulated speech with exaggerated intonation, when compared to normal speech. Several suggestions for the role of these properties have been proposed, and the sustaining idea is that IDS may help an infant to bootstrap its language skills, especially in creation of native language specific phonetic categories and boundary detection in word segmentation (Thiessen et al., 2005).

Kuhl et al. (1997) showed that IDS leads to a stretching of the formant space, increasing the mean distance between /a/, /i/ and /u/ vowels in terms of $F_1/F_2$ coordinates. In other words, they found that phonetic units were more clearly separated from each other in IDS, which may aid learning of the units. Enlargement of the vowel triangle also increases the phonetic variability (de Boer and Kuhl, 2003), which has been shown to aid in category acquisition (Lively et al., 1993; Kuhl et al., 1997). It has also been suggested that the increased separation of phone classes occurs during stressed (pitch-accented) vowels that are common in IDS speech, while unstressed or unaccented vowels show more coarticulatory reduction, leading to shrinking of the phonetic space (Kirchhoff and Bilmes, 1999; Sluijter and van Heuven, 1996). Van de Weijer (2001) also found that expansion of the vowel space in IDS occurred during content words, while in surrounding function words the effect was reversed.

The use of IDS has also been investigated in automatic speech recognition (ASR) by Kirchhoff and Schimmel (2005) in order to determine whether the use of IDS would reduce the size of training material needed to train HMM-models. However, they did not observe any direct advantage for using IDS instead of ADS in training, although recognizers trained with IDS had a relatively lower loss of performance with ADS data than ADS trained recognizer on IDS data. Closer analysis revealed that IDS had larger phone-class overlaps in feature space, which is an interesting finding, since it partially contradicts with other findings that promote the notion of more clearly segregated phone classes in infant-directed speech (e.g., Kuhl et al., 1997).

The purpose of this paper is to review the findings from experiments that compared properties of infant-directed speech (IDS) and adult-directed speech

(ADS) spoken in both Finnish and Swedish. The material was recorded as a part of a multilingual ADS/IDS corpus that was designed for ecologically and cognitively plausible training of a language-learning agent. The first section of this paper describes the spectral properties of ADS and IDS analyzed by a method that utilizes bottom-up clustering of phone-like speech segments extracted from speech (see Räsänen, 2007, for an overview of the algorithm). The second section presents findings from pitch and tilt analyses of IDS and ADS. Finally, the third section covers findings regarding segmental durations in these different speech types.

## *2 Material*

Speech material was recorded as a part of the ACORNS project, in which three corpora of different linguistic complexity are in the process of being created for Finnish, Swedish, Dutch, and English with gradually increasing grammatical, lexical, semantic and phonetic complexity. The material used in these experiments was taken from the first year corpus for Finnish and Swedish and contains simple utterances with a limited amount of different sentence structures and words. The material was spoken by eight different speakers, of which four speakers were native Finnish and four speakers were native Swedish, two males and two females for each language. Two thousand utterances were recorded for each speaker, one thousand utterances of adult-directed and one thousand utterances of infant-directed speech. The linguistic content of these IDS and ADS sets was identical.

Sentences were constructed using simple rules to combine a carrier sentence with a keyword. In general, a total of 10 carrier sentences, such as “Here is the X”, “Can you see the X?”, etc. where X was one of the 10 possible keywords, e.g., “car”, “ball” “telephone”, etc. These combinations led to 100 unique sentences being formed. Each sentence was then prompted 10 times in ADS and 10 times in IDS, to capture naturally occurring variations, yielding a total of 2000 prompts. During the recording procedure a speaker was shown one prompt at a time in sets of 10 prompts for ADS, and then followed by the same 10 prompts in IDS. This frequent switching between ADS and IDS was used to retain speaking style contrast. The required speaking style for a prompt was accomplished by visually indicating on the prompt screen an image of their spouse (to elicit ADS), or an image of their infant (to elicit IDS).

## *3 Experiment 1: Classification by spectral properties*

This experiment aimed to investigate spectral differences between ADS and IDS in order to find out if such differences exist and whether they can be utilized to differentiate the two speech types. The analysis was carried out by examining the behavior of incremental spectral clustering of the segments.

### *3.1 The algorithm*

The blind segmentation algorithm that is currently under development uses a Mel-frequency cepstral coefficient representation of a signal for the detection of segment boundaries. The signal is first windowed into short 6 ms frames and the cross-correlation of frames is utilized to enable detecting sudden changes in the signal. Segment boundaries are hypothesized to exist at locations where the change in the spectral properties exceeds a manually defined threshold level. For feature extraction, mean energy is removed from the MFCC frames and the vectors are normalized to zero mean unit vectors ( $\sum c_i^2 = 1$, $N = 12$ coefficients). Each segment is divided into an onset and offset section (initial 40 % and final 60 % of the duration) and the algorithm picks the five most contrastive MFCC vectors for both sections and averages them into two representations for each segment. Vector contrast for a frame is defined by the standard deviation of time-domain amplitude in a window from which MFCC coefficients are computed. Segments are classified into two categories based on their mean energy level: all segments with mean energy between $E_{min}$ to $E_{min}$ + 7 dB, where $E_{min}$ is the minimum energy of the signal, become assigned to the *background/silence* category. All segments exceeding the background noise signal energy level by 7 dB or more are assigned to the *main* category.

The onset and offset sections of the segments are clustered into separate spaces by using a simple incremental algorithm that computes the cross-correlation of the incoming segment's spectral vector to all existing clusters. An input is merged to the best matching cluster if the merging threshold $d_{min}$ is exceeded. If no suitably close matching cluster is found, a new cluster is created. Clusters retain information only about their centroid, i.e., the mean spectrum of all merged segments. Therefore, four combinations of spaces exist in total: onset/ offset (2) x energy (2).

The clustering algorithm described above is similar to the well-known *k*-means approach, but instead of requiring the entire classified material at once, it works on a segment-by-segment basis. The number of clusters is not known beforehand. This leads to a cognitively more plausible process, where the signal entering the system is first converted into a feature description, and the features representing the entire utterance are stored for a maximum of only a few seconds. Segmental features are used to activate and update categorical "phone-centers" in a multi-dimensional space but are not stored separately. Ultimately, complex acoustic waveform information is compressed into sequences of cluster labels without external knowledge interfering with the process, an approach that combines many aspects of learning vector quantization (LVQ; see, e.g., ) and self-organizing maps (Kohonen, 1984). Naturally, the classical k-means approach would provide better classification accuracy in terms of cluster selectivity if the number of clusters is chosen wisely, since access to the entire group of classifiable

items is available at all times. A more comprehensive description of the segmentation and clustering algorithm, including performance evaluations, can be found in Räsänen (2007).

### *3.2 Results*

The material was strictly divided into ADS and IDS content. Clustering was performed to create four different models for each speaker: an IDS-trained model and an ADS-trained model, both with two different merging (spectral similarity) thresholds $d_{min} = 0.5$ and $d_{min} = 0.7$. These sets of cluster spaces can be considered as IDS and ADS models with lower and higher selectivity, denoted as $MI_{0.5}$, $MI_{0.7}$, $MA_{0.5}$ and $MA_{0.7,}$ respectively.

The first finding was that the average number of clusters in IDS-spaces was larger than in ADS-spaces with both thresholds (+6.5 % with $d_{min} = 0.5$ and +10.7 % with $d_{min} = 0.7$), although some speakers did not exhibit significant differences. This is an expected result, as IDS is often considered as more vividly expressed than ADS and therefore some prosodical aspects (e.g., clearer articulation and more stress) may expand the formant space, producing more variations (Kuhl et al., 1997).

A more variable pitch may also cause changes in spectral representations as it affects the density of the formant structure. One explanatory factor may also be the MFCC window size: a 10 ms window may or may not capture properties of F0 depending on the frequency of the pitch. If pitch becomes relatively high (which is characteristic for IDS, see experiment 2), it will reveal itself in the spectral representation (which may not happen in the ADS case). Since cluster representations contain integrated properties of tens or hundreds of segments (both in terms of the Mel-band integration of the frequency domain and integration of several such representations together), pitch periodicity itself was not visible in the post-analysis of spectral properties to verify this assumption.

Also, an interesting notion is that the Swedish speech, consisting of similar simple structured utterances and with the same amount of different carrier sentences and keywords as in Finnish, yields on average of 52 % more clusters for both ADS and IDS. There may be several explanations for this difference, but one might be that since Swedish is a stress-timed language where most of the stressed words are accentuated, and this leads to variable $F_0$ modulation contours (Fant et al, 2000; Gårdning, 1989; Bruce and Granström, 1993) and therefore to the possibly spectral representation effects noted in the previous paragraph. On the other hand, Finnish is a syllable-timed language and may not exhibit this type of behavior to such a degree.

Clustering ADS to an IDS-model and then comparing the results to IDS clustered to the same model, tested the mutual compatibility of ADS and IDS models. Similarly, IDS was clustered to an ADS model and compared to ADS.

The mean of correlations between each segment and the best matching cluster was computed for both cases. As expected, IDS utterances were slightly better matched with IDS clusters than with ADS clusters (e.g., difference $\Delta_{corr} = 0.012$, with $d_{min} = 0.7$) and ADS utterances better with ADS clusters than IDS clusters ($\Delta_{corr} = 0.0071$). The overall differences were not very large but systematic, as all speakers had better correlation in congruent conditions with the higher threshold (there were two speakers with non-existent differences for the congruent and non-congruent situation with the $d_{min} = 0.5$ models). On average, ADS speech was slightly more efficiently mapped to the IDS space than IDS speech to the ADS space ($\Delta_{corr} = 0.005$) but the difference is not statistically significant. Nevertheless, this finding is in line with the work of Kirchoff and Schimmel (2005), who observed that an ASR (HMM) system trained with IDS and tested with ADS had relatively lower performance loss when compared to a system trained with ADS and tested with IDS.

The final experiment with the clustering algorithm was to find out whether cluster space models for infant-directed speech (MIDS) and adult-directed speech (MADS) were actually able to detect whether the input speech was directed to infants or adults despite the relatively small differences in adult- and infant-directed models. Each utterance was clustered simultaneously to MIDS and MADS and the model producing the better average match for all segments in the utterance was considered to indicate who is being addressed. Figure 1 shows the mean recognition rates for IDS and ADS utterances for all eight speakers. As can be seen, the recognition rate is significantly above chance with all speakers but accuracy also varies greatly between speakers. Recognition rates for specific speakers also correlate well with perceptual impressions from listening to ADS and IDS utterances: speakers with high recognition rates have a more lively and dynamic IDS compared to their ADS, whereas it may be difficult to detect perceptual differences between ADS and IDS in cases of low recognition rate speakers. Swedish recognition rates are also on average 10 % higher than in Finnish. Speakers #1, #2, #5 and #7 in Figure 1 are female, and no gender specific trend in recognition can be detected.

Figure 2 demonstrates the distributions of the difference between the correlation to the IDS trained space and ADS trained space for both ADS and IDS material for Speaker #5 (i.e., the criterion for deciding whether speech is IDS or ADS). In this case the mean recognition rate of the speaker is well above 90 %, and therefore the distributions' centers of masses are well separated from each other.

All things considered, spectral variability of IDS compared to ADS depends greatly on the individual differences in speaking style. For some speakers the spectral variability of speech segments is larger between the two styles, although all speakers exhibit small biases in the structure of the phone-space as the recognition rates in ADS- or IDS-decision task are above chance.

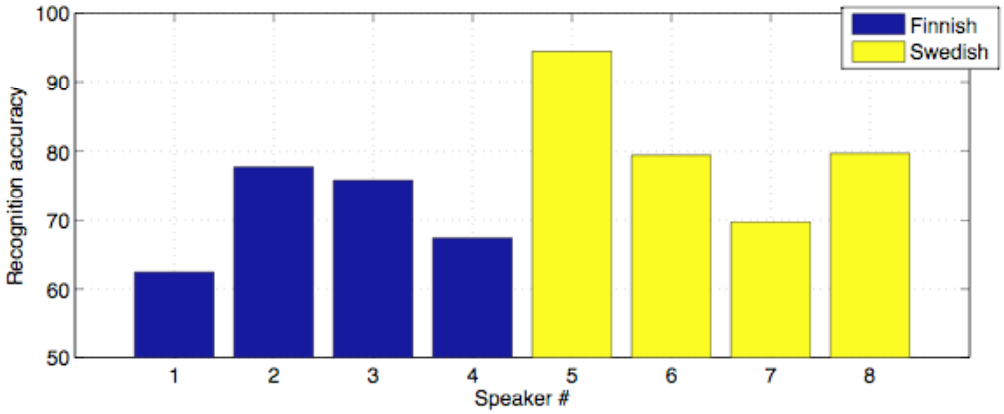


Figure 1. Mean differentiation ability of IDS and ADS for different speakers (50 % = chance level).

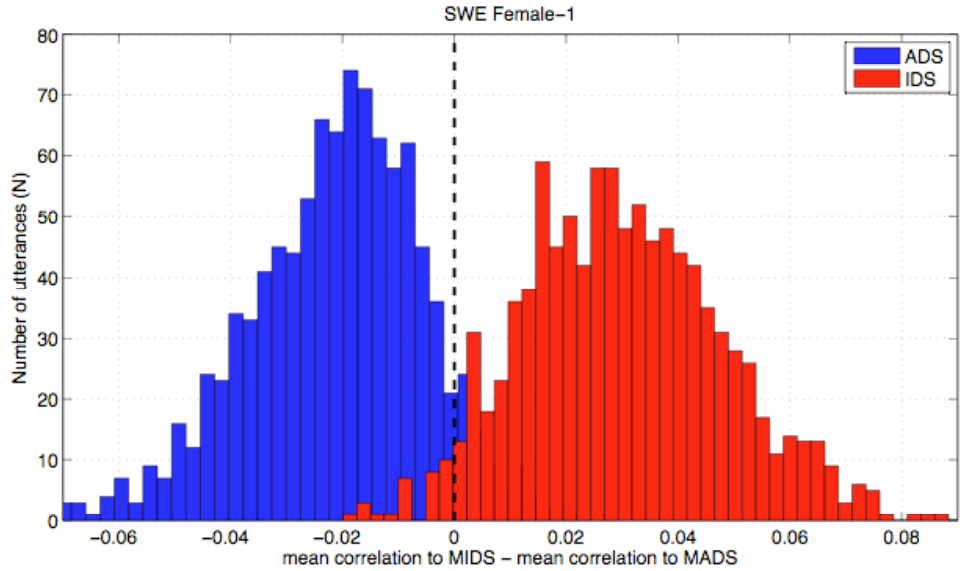


Figure 2. Distributions of correlation distance difference between models MIDS and MADS for ADS and IDS material, Swedish female speaker (Speaker #5 in Figure 1).

## *4 Experiment 2: Pitch and tilt analysis*

Pitch is a cue to several levels of linguistic information (Fisher and Tokura, 1996) and it has been determined that pitch contours are exaggerated in infant directed speech (Stern et al., 1983; Papousek and Hwang, 1991; Fernald and Simon, 1984; Jacobson et al., 1983), including tonal languages (Grieser and Kuhl, 1988). It has

been suggested that this may help infants to detect unit boundaries in acoustic patterns (Fisher and Tokura, 1996), promote infant attention and signal positive affect (Grieser and Kuhl, 1988). We tested pitch behavior in ADS and IDS using ACORNS corpus material hoping to replicate these findings.

### *4.1 Pitch analysis*

Distributions of pitch ($F_0$) were analyzed for ADS/IDS differences. Signals were windowed using a 40 ms Hamming-window with a 5 ms step and standard cepstral analysis performed. Frames without a sufficient $F_0$-peak between 50 and 600 Hz were discarded leading effectively to only the analysis of voiced segments of speech. F0 values from all frames for all ADS and IDS material from each speaker were stored and histogram analysis performed. Figure 3 shows pitch histograms for four speakers.

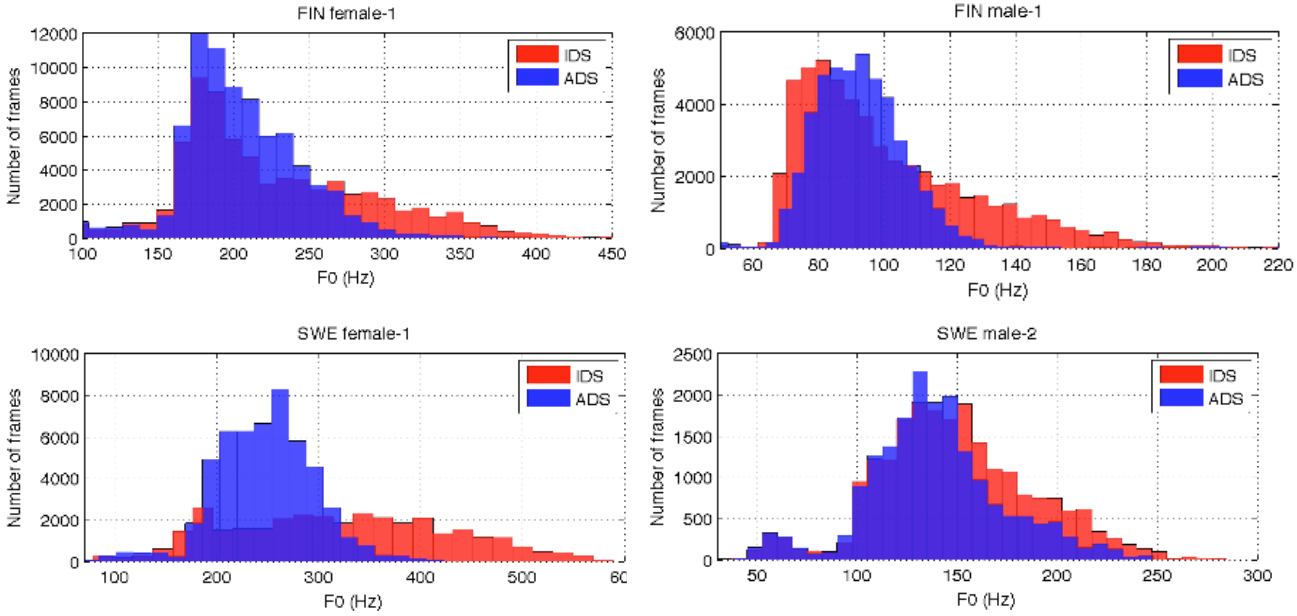


Figure 3. Histograms of F0-values for two females (left) and two males (right), one gender per language.

As expected, the difference in $F_0$ distributions between ADS and IDS is noticeable for all but one speaker: IDS contains a broader distribution of $F_0$-values, creating a tail area for IDS distributions at higher frequencies. Also, the size of the difference between ADS and IDS distributions correlates well with subjective perceptual judgments regarding the liveliness of the speech. The one exception is a Swedish male speaker whose ADS and IDS have very similar distributions and nearly identical means despite the fact that the speaker utters ADS sentences with

a greater tempo and his IDS is systematically much more carefully articulated and stressed.

As a conclusion, the findings of $F_0$ analysis confirm the previously established view found in the literature that $F_0$ is more variable in infant-directed speech.

### *4.2 Tilt analysis*

Spectral tilt of the speech material was studied in order to determine whether infant-directed speech differs from adult-directed speech in this manner. Since IDS and ADS speech material was identical in terms of linguistic content of spoken utterances, i.e., there was a corresponding pair for each IDS utterance in ADS, the spectral tilt of the utterances was examined pair-wise to elaborate the statistical salience of the differences between the two.

Both signals in the pair were windowed using a 45 ms Hamming-window with a 10 ms step size. The abs-FFT spectrum was then computed for each frame *k*. Tilt $a_1$ of the spectrum was estimated by fitting a line to the data in a least squares sense and the slope for each frame was stored for analysis. The mean tilt and variance of tilt were computed in parallel for both utterances and their ratios $t_r$ and $var_r$ (eq. 1 and 2) determined. These values were computed for all 1000 pairs per speaker to obtain speaker specific results.

$$t_r = \frac{\frac{1}{N_i}\sum_{k=1}^{Ni} a_{1,k,IDS}}{\frac{1}{N_a}\sum_{k=1}^{Na} a_{1,k,ADS}} \tag{1}$$

$$\mathrm{var}_r = \frac{\sum_{k=1}^{Ni} (a_{1,k,IDS} - \mu_{IDS})^2}{\sum_{k=1}^{Na} (a_{1,k,ADS} - \mu_{ADS})^2} \tag{2}$$

With this small amount of speakers no systematic difference in tilt could be detected (Table 1). Although the two Finnish male speakers exhibit overall steeper tilts and greater tilt variances over IDS compared to ADS utterances than their female counterparts, this effect cannot be seen in Swedish speakers. Also, since the standard deviation of tilt and variance ratios between utterance pairs is large and the number of speakers in the material is small, it is difficult to make any far-reaching conclusions in this case. It should also be noted that the rate of speech affects the tilt means when computed over entire utterances and therefore might cause biasing in the IDS versus ADS condition.

Table 1. Tilt ratios IDS/ADS and their standard deviation for all speakers.

| Swedish | $t_r$ | σ | $var_r$ | σ |
|---|---|---|---|---|
| **Male 1** | 0.882 | 0.153 | 0.838 | 0.364 |
| **Male 2** | 0.915 | 0.203 | 1.154 | 0.664 |
| **Female 1** | 0.737 | 0.224 | 0.730 | 0.603 |
| **Female 2** | 1.052 | 0.202 | 1.239 | 0.730 |
| **Finnish** | | | | |
| **Male 1** | 1.256 | 0.251 | 1.444 | 0.598 |
| **Male 2** | 1.076 | 0.176 | 1.263 | 0.450 |
| **Female 1** | 0.911 | 0.165 | 0.875 | 0.345 |
| **Female 2** | 0.793 | 0.232 | 0.669 | 0.417 |

In addition to examining entire signals, tilt analysis was repeated for only voiced frames of speech (voicing detected with standard cepstral analysis). This did not significantly affect the findings above.

The overall conclusion from tilt analysis is that there seems to be speaker specific properties in spectral tilting that differ systematically between adult- and infant-directed speech, but this systematic behavior does not extend to gender or language specific statistics.

## *5 Experiment 3: Segmental duration analysis*

This section investigates some of the effects that ADS and IDS exhibit on phone segmental duration. The material used in this analysis includes all of the Swedish and Finnish speech existing in the ACORNS Year 1 corpus from eight speakers. Since 2000 utterances were obtained from each speaker, a total of 8000 utterances existed per language that included 98488 Swedish and 88579 Finnish phones. Phones that were excluded from the duration analysis were silences and/or pauses at either the beginning or end of utterances, or pauses that existed between words. However, silences and pauses that occurred within words were included in the analyses.

Figure 4 shows an example of the Swedish sentence “Det är en bil” that was forced aligned using an automatically generated phonemic transcription and an HTK-based aligner. In this case the utterance was pronounced as [D E: M # B I: L] using the STA phonetic alphabet (Salvi, 2008). On the phone level the long pause between /det är en/ and /bil/, labeled with [#] and highlighted in green, is an example of a phone unit that is not included in the analysis. Furthermore, the single pause before the start of the utterance, as well as the two pauses after the utterance are also not included. However, the short pause within the word [B I: L] is included since it can be seen to contribute to the micro-rhythm of the sentence.

Phones that were deemed acceptable for use in further analyses were called "acceptable units".

It should be noted that the accuracy of the forced alignment for Swedish is in certain cases questionable since often high levels of reduction occurred, i.e., the canonical form of the orthography and the actual spoken phonetic form differed substantially. This effect can also be seen in Figure 4 since the phones [D E AE3 R E N] are being forced to model a section of speech that has been reduced to [D E: M # B I: L].

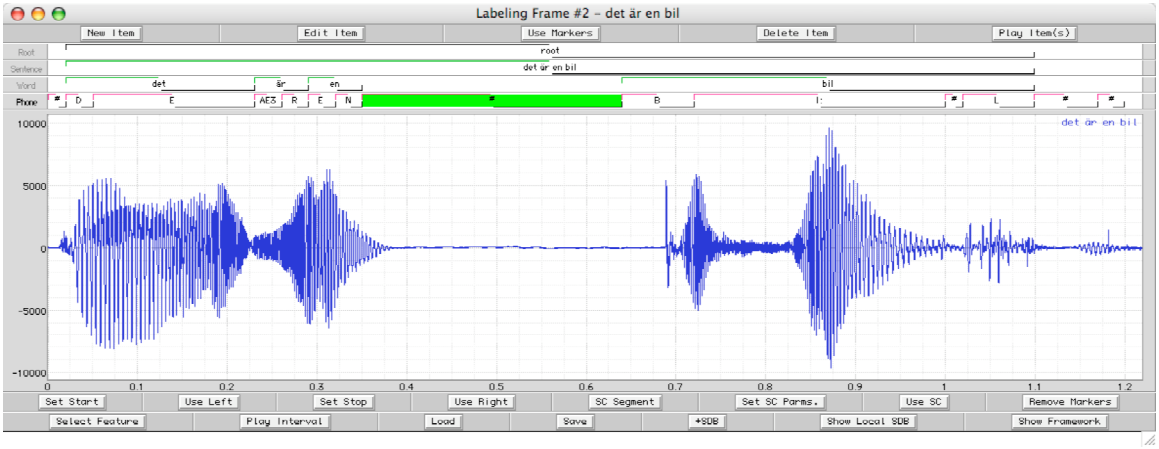


Figure 4. Forced alignment for the utterance /Det är en bil/ spoken by a female Swedish speaker. Phone tier labeled using STA.

Distributions of phone segment durations were then calculated using four different search queries to study the effects of ADS and IDS speaking styles. The first query searched for all acceptable phone units from the ADS speech of each speaker and the resulting set was called "All ADS phones". Likewise, all existing IDS phones for a speaker were queried for and named "All IDS phones". Since each utterance included an object or "focus" word, e.g., "car" in the sentence "This is a car", two more queries and phone set results were created to see if any change in duration existed within ADS or IDS phones existing within a focus word. These sets were called "Only ADS focus word phones" and "Only IDS focus word phones", respectively. Figure 5 shows in graphical form the average duration of each phone set (time (ms)) plotted against each of the four queries. As can be seen, larger variations in average duration occur for the four Swedish speakers (Anna, Björn, Nancy, and Olov) when compared to the other speakers who spoke Finnish.

Relative lengthening of all phones due to IDS being spoken instead of ADS varied from speaker to speaker, and is shown in Figure 6. The ratios for mean IDS vs. mean ADS for all phones ranged from a maximum value of 1.42 for male Swedish speaker Olov to a minimum value of 0.99 for Finnish speaker Seppo.

The latter indicates that IDS speech for this speaker was actually spoken more quickly than ADS. In general, the average lengthening of all phones for Swedish IDS was 24.7 % over the values calculated for ADS while for Finnish the same respective value was only 7.3 %. Average lengthening for both languages combined was 16 %. Figure 6 also indicates similar ratios but calculated over the phones existing within focus words only, i.e., the second column in the figure with the key "Focus IDS/ADS ratio". In general, phones uttered within IDS focus words do not receive notably more nor less amounts of lengthening when compared to their ADS counterparts and any smaller variation may be due to speaker specific behavior.

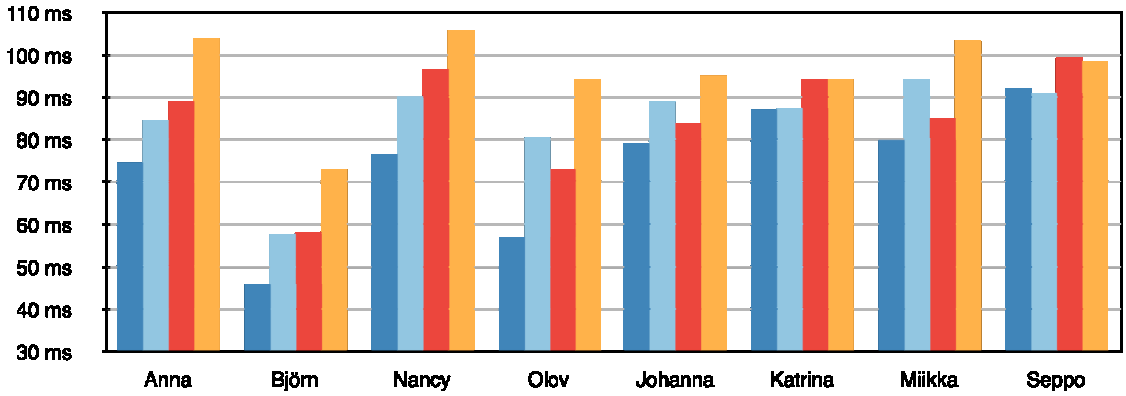


Figure 5. Average phone segment duration vs. speaker. A much larger range of variation in duration can be seen for the first four Swedish speakers as compared to the last four Finnish speakers.

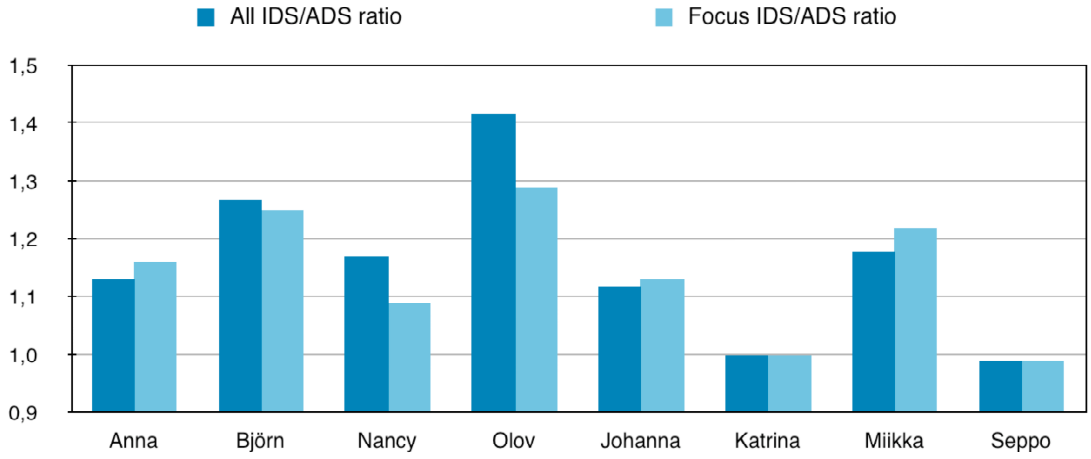


Figure 6. Relative lengthening of phone segmental durations for each speaker due to IDS being spoken instead of ADS. The left bar for each speaker represents the calculated ratio when "All IDS phones" is divided by "All ADS phones" for each speaker. The second bar represents the ratio when only phone durations existing in focus words are compared.

## *6 Discussion*

The most obvious differences between infant-directed and adult-directed speech types were found in pitch and duration. Distribution of pitch was found to be wider in IDS, which agrees with the literature that covers infant-directed speech. On average, IDS phones were found to be relatively longer than their ADS counterparts, but no systematical difference was detected between focus and non-focus words. Analysis of spectral tilt did not reveal any bias in favor of either speech types.

The number of the clusters in the cluster space, that is, variability of spectral descriptions of phone-like segments, was approximately 50 % larger in Swedish when compared to Finnish. No other reliable and systematic differences between Finnish and Swedish languages or genders were detected regarding the aspects investigated in this research.

***Acknowledgements***

The research was conducted as a part of ACORNS (Acquisition of Communication and Recognition Skills), an EU-funded Sixth Framework Programme, Priority [2], Future and Emerging Technologies, project, 2007–2009.