4 Discussion

Referring back to the issue about necessary and sufficient cues for word accent identification (cf. Introduction), we will comment on a number of points in the light of our experiment.

Is the gliding rise through the stressed syllable necessary for the perception of accent II? – Replacing this glide by an F0 jump up to the stressed syllable results in a sizeable decrease in the votes for accent II (cf. glide/dip4 vs. jump/dip4). This suggests that the gliding rise is necessary for the unambiguous perception of accent II.

Is a late-timed fall necessary for the perception of accent II? – Replacing the F0 fall through the post-stress syllable by a high plateau in the target word yields a tendency towards accent I (cf. glide/dip4 vs. glide/dip0). This suggests that the fall is necessary. However, the fall must not be substancially earlier than in the original accent II word, since this would correspond to accent I.

Thus, both the gliding rise and the late fall seem necessary for the unambiguous perception of accent II. When one of them is removed, the ratings tend towards accent I. When both these cues are absent, the tendency becomes rather strong (cf. jump/dip0).

What is necessary, and what is sufficient for the perception of accent I? – Accent I is most convincingly represented by stimuli with an early fall (dip5). However, the discussion above has already shown that this early fall cannot be a necessary cue, since a number of stimuli lacking this early fall have received high accent I ratings (cf. also jump/dip1-2). Furthermore, a high-starting stressed syllable (jump) favors accent I ratings, but cannot be regarded as sufficient, since the absence of an accent II-like fall appears necessary (cf. dip3).

Thus, our results are most easily explainable along the hypothesis that there are no specific necessary cues for accent I at all, but that simply the absence of accent II cues is sufficient for the perception of accent I. It is still remarkable, though, that the absence of only one accent II cue alone (e.g. the late fall) results in more votes for accent I than for accent II.

Why does a glide followed by a plateau trigger a considerable number of votes for accent I? – We do not have a definite answer to this question. One possibility is that the conditions of phrase intonation play a role. In the early part of a phrase, the expectation is a rising pattern. Thus a phrase-initial accent I may be realized as a rising glide even in South Swedish, as long as there is no immediately following F0 fall. Another possibility is that the glide-plateau gesture represents a typical accent I pattern of another dialect type (Svea or Stockholm Swedish), even if the context word at the end has a stable South Swedish pattern.

What does our experiment tell us about the markedness issue? – From the perspective of perceptual cues, accent II in South Swedish appears to be more "special" than accent I. This will lend some support to the traditional view of accent II being the marked member of the opposition (cf. Elert, 1964; Engstrand,1995; Riad 1998).

Acknowledgements

Joost van de Weijer assisted us with advice concerning methodology and statistics.

References

Bruce, G. & E. Gårding, 1978. A prosodic typology for Swedish dialects. In E. Gårding et al. (eds.), *Nordic Prosody*. Lund: Lund University, Department of Linguistics, 219-228.

Elert, C-C., 1964. *Phonologic Studies of Quantity in Swedish*. Stockholm: Almqvist & Wiksell.

Engstrand, O., 1995. Phonetic interpretation of the word accent contrast in Swedish. *Phonetica 52*, 171-179.

Riad, T., 1998. Towards a Scandinavian accent typology. In W. Kehrein & R. Wiese (eds.), *Phonology and Morphology of the Germanic Languages*. Tübingen: Niemeyer, 77-109.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 9–12

Focal Accent and Facial Movements in Expressive Speech

Jonas Beskow, Björn Granström, and David House
Dept. of Speech, Music and Hearing, Centre for Speech Technology (CTT), KTH, Stockholm
{beskow|bjorn|davidh}@speech.kth.se

Abstract

In this paper, we present measurements of visual, facial parameters obtained from a speech corpus consisting of short, read utterances in which focal accent was systematically varied. The utterances were recorded in a variety of expressive modes including Certain, Confirming, Questioning, Uncertain, Happy, Angry and Neutral. Results showed that in all expressive modes, words with focal accent are accompanied by a greater variation of the facial parameters than are words in non-focal positions. Moreover, interesting differences between the expressions in terms of different parameters were found.

1 Introduction

Much prosodic information related to prominence and phrasing, as well as communicative information such as signals for feedback, turn-taking, emotions and attitudes can be conveyed by, for example, nodding of the head, raising and shaping of the eyebrows, eye movements and blinks. We have been attempting to model such gestures in a visual speech synthesis system, not only because they may transmit important non-verbal information, but also because they make the face look alive.

In earlier work, we have concentrated on introducing eyebrow movement (raising and lowering) and head movement (nodding) to an animated talking agent. Lip configuration and eye aperture are two additional parameters that we have experimented with. Much of this work has been done by hand-manipulation of parametric synthesis and evaluated using perception test paradigms. We have explored three functions of prosody, namely prominence, feedback and interrogative mode useful in e.g. multimodal spoken dialogue systems (Granström, House & Beskow, 2002).

This type of experimentation and evaluation has established the perceptual importance of eyebrow and head movement cues for prominence and feedback. These experiments do not, however, provide us with quantifiable data on the exact timing or amplitude of such movements used by human speakers. Nor do they give us information on the variability of the movements in communicative situations. This kind of information is important if we are to be able to implement realistic facial gestures and head movements in our animated agents. In this paper we will report on methods for the acquisition of visual and acoustic data, and present measurement results obtained from a speech corpus in which focal accent was systematically varied in a variety of expressive modes.

2 Data collection and corpus

We wanted to be able to obtain articulatory data as well as other facial movements at the same time, and it was crucial that the accuracy in the measurements was good enough for

resynthesis of an animated head. The opto-electronic motion tracking system, the Qualysis MacReflex system, that we use has an accuracy better than 1 mm with a temporal resolution of 60 Hz. The data acquisition and processing is similar to earlier facial measurements carried out at CTT by e.g. Beskow et al. (2003). The set-up can be seen in Fig. 1, left picture.

Figure 1. Data collection setup with video and IR-cameras, microphone and a screen for prompts (left) and a test subject with the IR-reflecting markers (right).

The subject could either pronounce sentences presented on the screen outside the window or be engaged in a (structured) dialogue with another person as shown in the figure. By attaching infrared (IR) reflecting markers to the subject's face (see Fig. 1), the system is able to register the 3D coordinates for each marker. We used a number of markers to register lip movements as well as other facial movements such as eyebrows, cheek and chin.

The speech material used for the present study consisted of 39 short, content neutral sentences such as *"Båten seglade förbi"* (The boat sailed by) and *"Grannen knackade på dörren"* (The neighbor knocked on the door), all with three content words which could each be focally accented. To elicit visual prosody in terms of prominence, these short sentences were recorded with varying focal accent position, usually on the subject, the verb and the object respectively, thus making a total of 117 sentences. The utterances were recorded in a variety of expressive modes including Certain, Confirming, Questioning, Uncertain, Happy, Angry and Neutral. This database is part of a larger database collected in the EU PF-Star project (Beskow et al., 2004).

3 Measurement procedure

In the present database a total of 29 IR-sensitive markers were attached to the speaker's face, of which 4 markers were used as reference markers (on the ears and on the forehead). The marker setup (as shown in Fig. 1) largely corresponds to the feature point (FP) configuration of the MPEG-4 facial animation standard.

In the present study, we chose to base our quantitative analysis of facial movement on the MPEG-4 Facial Animation Parameter (FAP) representation. Specifically, we chose a subset of 31 FAPs out of the 68 FAPs defined in the MPEG-4 standard, including only the ones that we were able to calculate directly from our measured point data.

We wanted to obtain a measure of how (in what FAPs) focus was realised by the recorded speaker for the different expressive modes. In an attempt to quantify this, we introduce the Focal Motion Quotient, FMQ, defined as the standard deviation of a FAP parameter taken over a word in focal position, divided by the average standard deviation of the same FAP in the same word in non-focal position. This quotient was then averaged over all sentence-triplets spoken with a given expressive mode.

4 Results and discussion

As a first step in the analysis, the FMQs for all the 31 measured FAPs were averaged across the 39 sentences. These data are displayed in Fig. 2 for the analyzed expressive modes, i.e. Angry, Happy, Confirming, Questioning, Certain, Uncertain and Neutral. As can be seen, the FMQ mean is always above one, irrespective of which facial movement, FAP, is studied. This means that a shift from a non-focal to a focal pronunciation on the average results in greater dynamics in all facial movements for all expressive modes. It should be noted that these are results from only one speaker and averages across the whole database. It is however conceivable that facial movements will at least reinforce the perception of focal accent. The mean FMQ taken over all expressive modes is 1.6. The expressive mode yielding the largest mean FMQ is Happy (1.9) followed by Confirming (1.7), while Questioning has the lowest mean FMQ value of 1.3. If we look at the individual parameters and the different expressive modes, some FMQs are significantly greater, especially for the Happy expression, up to 4 for parameter 34 "raise right mid eyebrow".

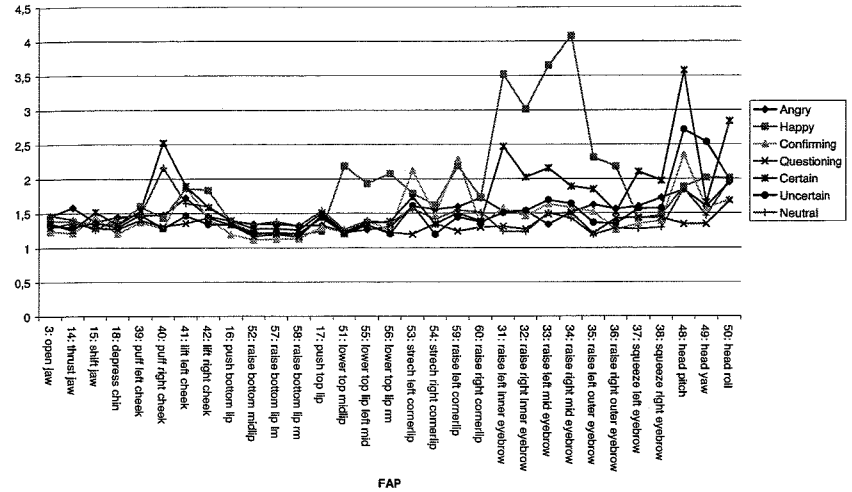

Figure 2. The focal motion quotient, FMQ, averaged across all sentences, for all measured MPEG-4 FAPs for several expressive modes (see text for definitions and details).

In order to more clearly see how different kinds of parameters affect the movement pattern, a grouping of the FAPs is made. In Fig. 3 the "Articulation" parameters are the ones primarily involved in the realization of speech sounds (the first 20 in Fig. 2). The "Smile" parameters are the 4 FAPs relating to the mouth corners. "Brows" correspond to the eight eyebrow parameters and "Head" are the three head movement parameters. The extent and type of greater facial movement related to focal accent clearly varies with the expressive mode. Especially for Happy, Certain and Uncertain, FMQs above 2 can be observed. The Smile group is clearly exploited in the Happy mode, but also in Confirming, which supports the finding in Granström, House & Swerts (2002) where Smile was the most prominent cue for confirming, positive feedback, referred to in the introduction. These results are also consistent with Nordstrand et al. (2004) which showed that lip corner displacement was more strongly influenced by utterance emotion than by individual vowel features.

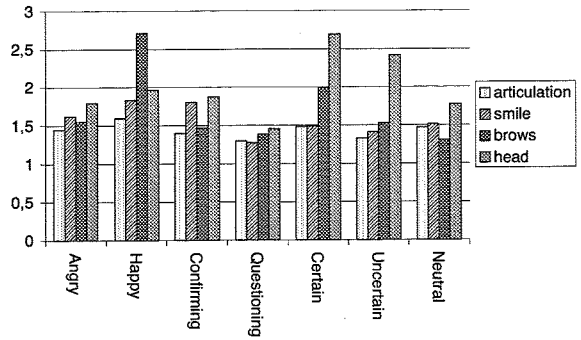

Figure 3. The effect of focus on the variation of several groups of MPG-4 FAP parameters, for different expressive modes

While much more detailed data on facial movement patterns is available in the database, we wanted to show the strong effects of focal accent on basically all facial movement patterns. Modelling the timing of the facial gestures and head movements relating to differences between focal and non-focal accent and to differences between expressive modes promises to be a fruitful area of future research.

Acknowledgements

This paper describes research in the CTT multimodal communication group including also Loredana Cerrato, Mikael Nordenberg, Magnus Nordstrand and Gunilla Svanfeldt which is gratefully acknowledged. Special thanks to Bertil Lyberg for making available the Qualisys Lab at Linköping University. The work was supported by the EU/IST projects SYNFACE, PF-Star and CHIL, and CTT, supported by VINNOVA, KTH and participating Swedish companies and organizations.

References

Beskow, J., L. Cerrato, B. Granström, D. House, M. Nordstrand & G. Svanfeldt, 2004. The Swedish PF-Star Multimodal Corpora. *Proc. LREC Workshop, Multimodal Corpora: Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces*, Lisbon, 34-37.

Beskow, J., O. Engwall & B. Granström, 2003. Resynthesis of Facial and Intraoral Articulation from Simultaneous Measurements. *Proc. ICPhS 2003*, Barcelona, 431-434.

Granström, B., D. House & J. Beskow, 2002. Speech and gestures for talking faces in conversational dialogue systems. In B. Granström, D. House & I. Karlsson (eds.), *Multimodality in language and speech systems*. Dordrecht: Kluwer Academic Publishers, 209-241.

Granström, B., D. House & M. Swerts, 2002. Multimodal feedback cues in human-machine interactions. *Proc. Speech Prosody 2002*, Aix-en-Provence, 347-350.

Nordstrand, M., G. Svanfeldt, B. Granström & D. House, 2004. Measurement of articulatory variation in expressive speech for a set of Swedish vowels. *Journal of Speech Communication 44*, 187-196.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 13–16

A Study of Simultaneous-masking and Pulsation-threshold Patterns of a Steady-state Synthetic Vowel: A Preliminary Report

Ulla Bjursäter
Department of Linguistics, Stockholm University
ullabj@ling.su.se

Abstract

This study will be a remake in part of Tyler & Lindblom "Preliminary study of simultaneous-masking and pulsation-threshold patterns of vowels" (1982), with the use of today's technology. A steady-state vowel as masker and pure tones as signals will be presented using simultaneous-masking (SM) and pulsation-threshold (PT) procedures in an adjustment method to collect the vowel masking pattern. Vowel intensity is changed in three steps of 15 dB. For SM, each 15 dB change is expected to result in about a 10-13-dB change in signal thresholds. For PT, the change in signal thresholds with vowel intensity is expected to be about 3-4 dB. These results would correspond with the results from the Tyler & Lindblom study. Depending on technology outcome, further experiments can be made, involving representations of dynamic stimuli like CV-transitions and diphthongs.

1 Introduction

This study is an attempt to partially replicate Tyler & Lindblom "Preliminary study of simultaneous-masking and pulsation-threshold patterns of vowels" (1982). Their intention was to investigate the effect of the two different masking types as well as the role of suppression in the coding of speech spectra.

Suppression, or lateral inhibition, refers to the reduction in the reaction to one stimulus by the introduction of a second (Oxenham & Plack, 1998). The ability of one tone to suppress the activity of another tone of adjacent frequency has been thoroughly documented in auditory physiology (Delgutte, 1990; Moore, 1978). In speech, suppression can be used to investigate formant frequencies.

In the original article, the authors (Tyler & Lindblom, 1982) constructed an experiment masking steady-state synthetic pure tones by simultaneous and pulsation-threshold patterns of vowels. Their vowels were synthesized on an OVE 1b speech synthesizer (Fant, 1960) with formant frequencies, bandwidths and intensities as approximate values for Swedish. In this study only one of the vowels from the original experiment is synthesized, using Madde, a singing synthesizer (<www.speech.kth.se/smptool/>) instead of OVE 1b.

In this experiment, the original vowel masking patterns will be used on the Swedish vowel /y/, a vowel that, according to Tyler & Lindblom (1982), is particularly useful in testing the role of suppression in speech as it has three closely spaced high frequency formants (F2, F3 and F4). F2 and F4 have about the same frequency as in the vowels /i/ and /e/, and a distinct perception of these three vowels must depend on good frequency resolution of F3 (Carlson et al., 1970; Bladon & Fant, 1978; Tyler & Lindblom, 1982).