PROSODY PROCESSING IN SPEECH PRODUCTION: PRE-EVALUATION OF A fMRI STUDY

Jörg Mayer
Experimental Phonetics, IMS, University of Stuttgart, Germany
joemayer@ims.uni-stuttgart.de

ABSTRACT

Due to noise emission of the MR scanner during functional imaging it is impossible to evaluate subjects' performance on speech production tasks online. Therefore we choose a pre-evaluation paradigm to extract the relevant phonetic parameters that characterize subjects' utterances. All experimental tasks were first recorded in an anechoic chamber before subjects passed on to the fMRI session. All subjects were recorded under two conditions: (1) silence and (2) with fMRI noise presented over headphones. We applied this procedure to answer two questions: (1) Does the scanner noise influence speech production, particularly the prosodic features of the speech signal? (2) Is the subjects' performance on experimental tasks as intended? Our results indicate that MR scanner noise does not significantly interact with prosody generation and that our experimental paradigms yield the intended results.

1. INTRODUCTION

1.1. Studying speech production with fMRI

At present, perception designs are predominant in functional Magnetic Resonance Imaging (fMRI) studies on speech and language processing. This is due to at least two problems which fMRI production studies have to cope with. The first problem are analysis artefacts in the functional images caused by movements of the articulators and the head. This difficulty has been solved by recourse to inner speech experiments or, more recently, by event-related fMRI techniques. The second problem is the evaluation of subjects' speech production. What do they utter while being scanned? Noise emission of the MR scanner during functional imaging makes online auditive evaluation impossible. Subsequent evaluation, though possible with noise-filtered speech signals, is only appropriate for the analysis of syntactic or semantic characteristics of utterances. However, if phonetic parameters of utterances are of central interest, filtering has essential shortcomings.

In our research on functional imaging of prosody processing in speech production [1] we want to find out whether subjects are able to produce the required prosodic variations and how they realize the different prosodic patterns. To examine these issues we chose a pre-evaluation paradigm as an alternative solution to the evaluation problem. All experimental tasks (see below for details) were first recorded in an anechoic chamber before subjects passed on to the fMRI session. The obtained high quality recordings were subjected to a detailed phonetic analysis and served as an estimation of the subjects' performance during functional imaging.

Besides the evaluation problem, another open question in studying speech production with fMRI is, whether the typical scanner noise influences verbal behavior during fMRI measurements. This question is of particular interest when the phonetic performance (rather than syntactic, semantic or other higher level processes) is subject to functional analysis. To examine noise effects we recorded each experimental task under two conditions: (1) silence (subjects hear only their own echo over headphones), and (2) with fMRI noise presented over headphones.

1.2. Studying prosody generation with fMRI

fMRI experiments require designs with task pairs where both tasks are identical except for the presence or absence of the one cognitive process that is examined — the *cognitive subtraction paradigm* (see [2] for an alternative approach). If prosody generation is the process in question, what is an appropriate baseline task?

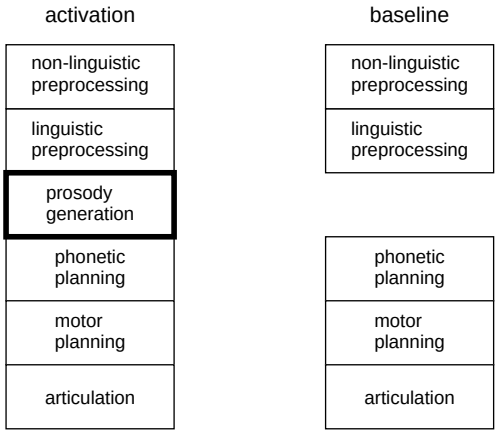

Figure 1. Schematic representation of cognitive processes involved in speech production and requirements of the cognitive subtraction paradigm.

In order to achieve parallel processing — except for prosody generation — below the linguistic preprocessing level we introduced a monotonous condition (baseline tasks) as opposed to the 'prosodic' condition (activation tasks). Similar items have to be produced in a monotonous manner, i.e., with invariant default settings of intensity, F0 and durational parameters, or with different prosodic patterns, i.e., varying settings of the relevant phonetic parameters. We assumed that under the prosodic condition the 'prosody generator' [3] is activated, whereas under the monotonous condition the production process proceeds identically except for the activation of the prosody generator.

But since prosodic utterances are the unmarked case in natural language production and since the output of higher-level linguistic processing must pass through the prosody generator to

reach phonetic planning (cf. [3]) — how can inhibitory activation in monotonous productions be avoided? For this purpose we used reiterant speech in all conditions of our experiment. Rendition of nonsense syllables enters the speech production process low enough to eliminate higher-level linguistic processing while subsequent processing levels are preserved. We assume that with reiterant speech, segmental spellout procedures are reduced to a minimum and that under the monotonous condition the output passes directly into the phonetic modules. Under the prosodic condition segmental spellout proceeds through the prosody generator, which is invoked with its full functionality (cf. [4]).

The following experiment was carried out to test the assumptions stated above and to evaluate subjects' performance on the sketched tasks.

2. METHODS

For the study 9 native German subjects were recruited (five females, four males, mean age 26.2 years, range 21-32 years) who were paid for the participation in the experiment. The subjects were asked to produce a sentence-like sequence consisting of five syllables [dadadadada] with various pitch-accent types and locations (the FOCUS condition), with various boundary tone types (the MODUS condition), and with various kinds of emotional state marking (the AFFECT condition). As baseline task they were asked to produce the sequences [dadadadada, dididididi, dododododo, dududududu] in a monotonous voice (with a syllable frequency of about 5 Hz). The material is summarized in table 1 (intonational annotation is in accordance with [5]). The different paradigms applied as follows: In part 1 of the experiment blocks of 4 tokens either of FOCUS or baseline condition alternated. In part 2 blocks of 4 tokens either of prosodic or baseline condition alternated; in the prosodic blocks MODUS and AFFECT tokens alternated. Both parts consisted of 8 blocks each, 4 under monotonous condition and 4 under prosodic condition. This results in a total of 32 monotonous tokens, 16 FOCUS tokens, 8 MODUS tokens, and 8 AFFECT tokens per subject. All renditions were elicited by visual stimuli. Each stimulus was assigned to one prosodic type and one monotonous type, depending on the block instruction (cf. table 1).

Recordings were made in an anechoic chamber with high quality equipment. To simulate the scanning situation subjects were asked to lie down in front of a screen, on which stimuli were presented. The entire experiment was carried out twice: (1) in si-

	Stimulus	Paradigm	
		1 (FOCUS)	4 (MONOTONOUS)
Part 1		dadada**da**da H*L	dididididi
		dadada**da**da H*L	dadadadada
		**da**dada**da**da H*L H*L	dududududu
		**da**dada**da**da L*H L*H	dododododo
		2 (MODUS)	4 (MONOTONOUS)
Part 2		dadada**da**da H*L L%	dididididi
		dadada**da**da L*H H%	dadadadada
		3 (AFFECT)	4 (MONOTONOUS)
		dadada**da**da H*L [HAPPY]	dududududu
		dadada**da**da H*L [SAD]	dododododo

Table 1. Visual presented stimuli and reaction paradigms.

lence (only echo was presented over headphones), and (2) with noise presented over headphones; noise was recorded earlier in the MR scanner.

Recordings were digitalized and processed using ESPS/waves. The following parameters were calculated: 'sentence' (s) duration, F0 mean, and F0 range and vowel (v) duration, F0 mean, F0 range, F0 standard deviation, RMS mean, and RMS max. F0 parameters were taken from median filtered fundamental frequency estimations (filter width: 5).

3. RESULTS

3.1. Noise effect

All subjects generally spoke louder under the noise condition compared to the silence condition (RMS_mean: noise > silence). This effect was highly significant for all subjects ($p \leq .001$). Additionally, most subjects (n = 7) raised fundamental frequency under the noise condition (F0_mean(s): noise > silence; $p \leq .05$). 3 subjects prolonged 'sentences' as well (Duration(s): noise> silence; $p \leq .05$). F0 range was not affected by noise.

The noise effect did not interact significantly with any of the other experimental factors. Furthermore, inconveniences of the noisy condition did not lead to an increased error rate (which was in general extremly low).

3.2. Prosody generation

Let us now consider whether the presented design is appropriate for the study of prosody generation, and whether subjects show the intended behavior (all group results are based on Post Hoc Tukey HSD Multiple Comparisons).

3.2.1. Monotony. Vowel duration did not significantly vary between syllables 1 to 4 in the monotonously spoken items. Only the last syllable was prolonged (Duration(v): 1,2,3,4 < 5; $p \leq .001$). Intensity was decreased in the last syllable without variance in the first 4 syllables (RMS_mean: 1,2,3,4 > 5; $p \leq .001$). Fundamental frequency range was wider at the sequences' edges and stable through syllables 2 to 4 (F0_range(v): 1,5 > 2,3,4; $p \leq .001$). Mean F0 (F0_mean(v)) was stable throughout the sequences. All together, variation of prosodic phonetic parameters was exclusively restricted to the edges of monotonous items. Within the sequence, at least, all parameters under consideration were stable and invariant.

3.2.2. Intonation. Qualitative inspection of accent placement, accent type and boundary type variation showed, that all subjects were able to correctly and naturally realize the required intonation pattern. Quantitative analysis yielded vowel duration as the best predictor for accented syllables (Dur(v): accented > non-accented; $p \leq .001$), followed by mean F0. In H*L-accented syllables mean F0 was higher ($p \leq .001$) whereas in L*H-accented syllables mean F0 was lower than in unaccented syllables ($p \leq .001$). Concerning general vowel intensity, both RMS parameters reached only weak significance. 7 subjects varied maximum intensity (RMS_max: acc. > non-acc.; $p \leq .05$) and only 6 subjects showed an effect on mean intensity (RMS_mean: acc. > non-acc.; $p \leq .05$). The standard deviation of F0 within the vowel was not significantly affected by accentuation.

3.2.3. Affect marking. Concerning prosodic affect marking, inter-subject variability was relativly high. This is not surprising since the instructions for this condition were not very restrictive. In view of the functional imaging study, our aim was simply to force *some* affective load on speech processing under this condition. As a consequence, subjects chose individual strategies to express the difference between 'happy' and 'sad' renditions. Common to all subjects was the use of global fundamental frequency height to differentiate affective modes. The general tendency was to lower sad and to raise happy utterances with respect to neutral renditions (F0_mean(s): sad < neutral, $p \leq .05$ and neutral < happy, $p \leq .001$). Most subjects (n = 7) significantly increased global F0 range in happy utterances (F0_range(s): sad, neutral< happy; $p \leq .001$). Also 7 subjects varied expressivenes of H*L-accents depending on affective mode (F0_mean(v): sad < neutral < happy; $p \leq .001$). Durational and intensity markers, though applied by all subjects, revealed no consistent pattern. To sum up, all subjects significantly differentiated affective modes, but in doing so fell back on individual strategies.

4. DISCUSSION

The main effect of the noise condition — increased overall intensity — is to be expected and can be interpreted as a compensation strategy. The fact that the noise effect did not interact with the other experimental factors under consideration leads us to the assumption that the poor conditions in the MR scanner did not influence prosody generation. We think that our results are encouraging regarding language and speech production studies with fMRI, though the unavoidable noise of the scanner may increase error rates when dealing with more complex linguistic tasks.

The analysis of monotonous speech confirms our earlier assumption that reiterant 'unprosodic' speech directly enters the phonetic spellout procedures, by-passing the prosody generator. The observed patterns like prolongation of the last syllable (final lengthening) and phrase-final reduced intensity are well known phonetic effects which arise at the edges of various phonetic units. This means that phonetic processing is obviously involved. On the other hand, the absence of edge effects which are due to the prosody generator, like for example final lowering of fundamental frequency (cf. [6]), shows that higher level processing is evidenly not involved.

The results we have obtained from prosodic utterances are in accordance with recent studies on German stress and accent realization. [7] shows that duration is the best predictor for syllable prominence in German. When accounting for intensity aspects of prominence spectral tilt (which was not considered in our study) is a more promising measure than general RMS changes [8,9], which explains the weak effects observed concerning RMS parameters. The fact that fundamental frequency movement (F0 standard deviation) is not significant in accented syllables is not surprising. Pitch accents which have no particular function such as signalling special emphasis are usually realized as step accents in German [10]. Fundamental frequency steps up (H*L) or down (L*H) to the accented syllable and starts to smoothly fall or rise towards the end of the syllable. Our findings for mean F0 confirm this pattern.

In summary, these results support our experimental design, especially our assumption that reiterant speech is suitable to ini-

tiate natural prosody generation processing. Furthermore, regarding affect marking, we can state that all subjects prosodically differentiated the affective modes, although — apart from global fundamental frequency height variation — a consistent pattern did not arise. Finally, all subjects attending this pre-study showed comparable and 'correct' patterns in both prosodic and monotonous conditions. Therefore we may assume that the same subjects behaved very similarly during fMRI measurement which they attended afterwards, and that the analysis of the functional images does in fact yield insights into prosody processing in the human brain.

ACKNOWLEDGMENTS

This research is being supported by the grant nr. DO/536/2-1 from the German Science Foundation.

REFERENCES

[1] Mayer, J., Dogil, G., Wildgruber, D., Riecker, A., Ackermann, H. and Grodd, W. 1999. Prosody in speech production. *Proceeding from the 14th International Conference of Phonetic Sciences*. San Francisco 1999.

[2] Price, C.J. and K.J. Friston 1996. Cognitive conjunction: A new approach to brain activation experiments. *NeuroImage*, 5, 261-270.

[3] Levelt, W. 1989. *Speaking: From Intention to articulation*. Cambridge: MIT Press.

[4] Liberman, M.Y. and L.A. Streeter 1978. Use of nonsense-syllable mimicry in the study of prosodic phenomena. *Journal of the Acoustical Society of America*, 68, 231-233.

[5] Mayer, J. 1995. Transcribing German intonation: The Stuttgart system. Technical Report, University of Stuttgart.

[6] Liberman, M.Y. and J. Pierrehumbert 1984. Intonational invariance under changes in pitch range and length. In Aronoff, M. and R.T. Oehrle (eds.), *Language sound structure*. Cambridge: MIT Press, 157-233.

[7] Rapp, S. 1995. Maschinelles Lernen von Aspekten des deutschen Wortakzents. *Working Papers PhonetikAIMS*, Vol. 2,, No. 2, University of Stuttgart, 147-240.

[8] Sluijter, A.M.C. and V.J. van Heuven 1996. Spetral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100, 2471-2485.

[9] Classen, K., Dogil, G., Jessen, M., Marasek, K. and Wokurek, W. 1998. Stimmqualität und Wortbetonung im Deutschen. *Linguistische Berichte*, 174, 202-245.

[10] Fery, C. 1993. *German intonational patterns*. Tübingen: Niemeyer.