A PERCEPTUAL STUDY FOR MODELLING SPEAKER-DEPENDENT INTONATION IN TTS AND DIALOG SYSTEMS

Joachim J. Mersdorf, Thomas Domhöver

Institute of Communication Acoustics, Ruhr-University 44780 Bochum, Germany

Tel. +49 234 700 2470, Fax. +49 234709 4165, E-mail: mersdorf@aea.ruhr-uni-bochum.de

ABSTRACT

In general, most of the developed prosody and intonation models were obtained from a statistical analysis of F0 curves and resynthesis by TTS. But there is yet another chance improving quality and naturalness: effective results can also be obtained by analysing the listeners' common sense about natural intonational behavior. Therefore, we use a digital process that generates signals representing only the melody of the original speech signal. Comprehensive listening experiments become possible to analyse and compare the perception of natural and synthetic intonation. Based on the results of some listening experiments a statistical analysis of the F0 curves was carried out, regarding that a speaker-individual intonation model needs more quantitative F0 information than traditional descriptions. The aim is an prosodical speaker-dependent model for synthetic speech and dialog systems. Furthermore, this flexible approach should not be limited to speaker-individual intonation.

1. INTRODUCTION

Although synthesis by time-domain concatenation uses bricks of natural speech, it is well-known that prosodical attributes of the original inventory-speaker are lost. Additional modelling of these parameters will certainly improve the perceived naturalness and acceptance of synthetic speech. Further, speaking style and prosodical speaker characteristics become increasingly important to assign synthetical output of several natural input-speakers in multi-user dialog systems (e.g. like the German Verbmobil). Additionally, we acquire knowledge about strategies and methods in voice conversion. First of all we need a comprehensive auditive observation of intonational time-units to find out perceptual relevant speaker-dependent intonation parameters in read speech. Previous studies have shown that subjects are able to extract and identify speakers purely by an acoustical representation of their intonational melody [6]. Countless speaker-dependent parameters are well-known from the enhancement of technical speech recognition, but nevertheless in most cases these significant attributes are not perceptually relevant. A first statistically significant parameter seems to be the mean value and variance of F0 curves. But shifting and adjusting the mean of F0 and textual duration to a uniform value for all the speakers by using the PSOLA-technique does not result in lower recognition rates in many cases [6]. So it can be assumed that there are more parameters than these globals. Further, we discovered that the final F0 interval at sentence or phrase boundaries has a significant value for each speaker. Other studies refer to a significance of final F0 values at the end of phrases or sentences [10]. In order to find other perceptual parameters and to observe the relevant moments and prosodical units of recognition, a comprehensive listening experiment was developed. Reaction time measurements of recognition times were carried out to examine the temporal structures and moments of speech melody represented by F0 contours that are relevant for identifying speakers by their intonation. Another aim was to reduce the amount of speech material for further auditory and statistical analysis.

2. INTONATIONAL SIGNAL PRESENTATION

Although there is a general consensus that improved prosody contributes to a higher quality in synthetic speech, there is not yet any accepted standard methodology for an assessment test involving naive subjects [1] to seperately evaluate the prosodic components of synthetic speech. An alternative perceptual approach, in contrast to analysis by synthesis, is presented here.

2.1 Method of Stimulus Generation

The basis of the method was motivated by the idea of *Reiterant Speech* [2]. A digital signal process was developed [4] to reduce the different attributes of speech quality inside a speech-signal to some prosodical properties. The aim was to concentrate the attention of the subjects only on the quality and attributes of intonation. The process removes all the segmental information, typically represented by some spectral signal parameters, such as the formant frequencies. The first step is the determination of pitch by a time-domain algorithm and the computation of the envelope from the speech signal. In the second step, on every computed pitch-mark a glottal excitation pulse, like in the CCITT-Recommendation

P.50, is inserted. Every glottal pulse is weighted by the pitch-synchronous short-term root mean square value representing the envelope of the speech signal. The third step is a filtering of the whole glottal pulse-train by an time-invariant FIR-bandpass-filter in the fundamental frequency domain (50-350 Hz).

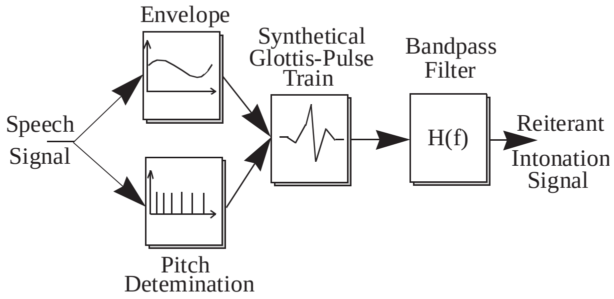

Fig. 1 Signal Process of *Reiterant* Intonation

The result is an intonation signal that sounds like speech, which is reduced to just about 6 % of the original spectral signal information. Subjects very often try to understand such a signal. In interviews they often say that the signals sound like speech but they miss the segmental intelligibility. In reality, the effect can be imitated by speaking during the lips are locked and fixed to the phoneme /*m*/ the whole time. Thus, the resulting signal can be considered as an acoustical correlation of speech intonation. Other approaches apply a saw-wave-train [5] or a steady amplitude synthetic /*a*/ [3]. In a first approach we now can concentrate the development, analysis and evaluation of intonation on using such signals. Later it can be replaced by PSOLA-manipulation of natural speech or by a speech synthesis system to evaluate the final model. Another important advantage of that process is, that intonation of natural and synthetic speech becomes more comparable, for example to quantify the distance of prosodical qualitiy of a TTS system and natural structures.

2.1 Method of Stimulus Presentation

An important criterion of participation in this listening experiment is that speakers from different speaking situations are well-known to all the subjects. 4 male speakers and 14 male subjects, all members of the institute, attended the experiment. The speakers' voices were familiar to the subjects from different speaking situations: lectures, presentations, discussions, and spontaneous dialogs. Three German texts from radio broadcasts [9] were read aloud by the four non-professional speakers, recorded and sampled in an unechoic chamber. Reading these texts aloud takes between 31-40 seconds. The test session contains 60 stimuli (4 speakers x 3 texts x 5 repetitions) in random choice. The subjects' task was to identify each speaker by his reiterant intonation signal as fast as they can. A software tool was developed, including a window with 4 buttons, representing the four speakers and one go-ahead-button in the center. Decisions and reactions were made by these buttons. First of all a profile of the 4 speakers was computed, representing some global parameters such as the mean value of F0 (113,5 - 134,5 Hz) and the standard deviation of F0 (17,2 - 31,2 Hz). The mean speech rate ranges between 3,8-4,9 syllables per second. Furthermore, distances in the mean of F0 between the 4 speakers were computed.

3. RESULTS OF AUDITORY EXPERIMENTS

	speaker1	speaker2	speaker3	speaker4
speaker1	**68,6 %**	5,7 %	16,2 %	9,5 %
speaker2	3,3 %	**75,2 %**	9,0 %	12,4 %
speaker3	13,3 %	12,4 %	**55,2 %**	19,0 %
speaker4	8,6 %	11,4 %	24,8 %	**55,2 %**

Tab.1: **recognition** and confusion results (presented speakers in rows, identified speakers in columns)

If we consider the theory in statistics having 4 alternative choices available, we can expect a recognition-rate of 25 %. But results are much higher: recognition-rates between 55,2 % - 75,2 % and lower confusion-rates between 3,3 % - 24,8 % were attained. The histogram *(Fig. 1)* representing the distribution density of reaction times shows an exponential decrease by time.

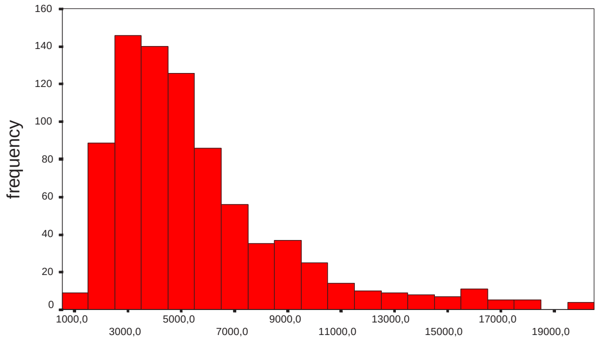

Fig. 2

The mean value for reaction-time is 6,0 seconds. All the correct reponses of reaction-times were now treated as recognition-time results. The boxplot in *Figure 2* gives an overview about the total recognition time for each speaker.

Speakers no. 1+2 were recognized faster and more frequently than speakers no. 3+4, because some parameters of the mentioned profile of speaker no. 3 and 4 seem to be very similar. In many cases speakers were already identified reliably (80-95% recognized) after the first or second sentence. The recognition-times are very similar for the three texts pesented; it depends on sentence structure and the number of syllables. In all cases less than 15 seconds were time enough to identfy the test speakers reliable only by intonational parameters. For read speech it can be assumed that the perceptual relevant parameters and differences already can be found in first sentences and phrases at the beginning of a text.

The resulting distribution function for each sentence and speaker very often contains terms of a higher increase in decision that often indicates moments of higher recogniton during the stimulus presentation.

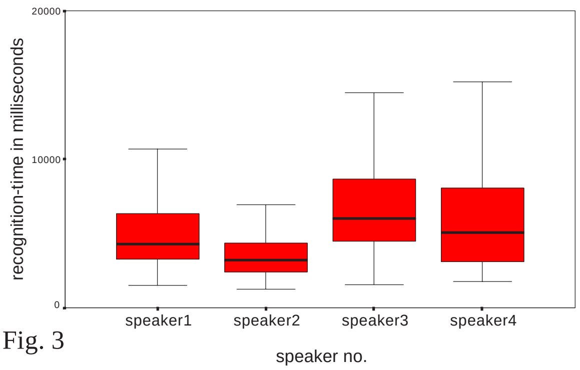

Fig. 3

4. RESULTS IN CONTOUR ANALYSIS

Having both a look at the contour and at the distribution of decisions, frequently an increase of recognition answers by time can be found in pauses after sentence and phrase boundaries, and very often after the fall of accents. Accepting a time-delay of reaction up to one second, we can assume that the fall interval (step-down) is important for human intonational recognition. Inspecting the contours we can notice that often decisions were made after H* L- tones, mostly after strong and sometimes after weak accents, after B3-boundaries and often during the following pauses. Analysing all local extreme F0 values results in typical fundamental frequencies for every speaker.

Adjacent F0 maxima and minima lead to a dynamical description by step-up and step-down intervals. This typical values can be found already very reliably in a single read text. *Figure 4* shows step-down intervals leading to some distinct minimum values, *Figure 5* shows typical step-up intervals (in semi-tones) relating to typical maximum values for each speaker.

Furthermore, virtual accents in unvoiced parts can be detected. Inspecting countless F0 contours we can often find an monotoneous increase part, interrupted by a short unvoiced part and followed by a decreasing of the fundamental frequency. For an automatic quantitative F0 analysis it could be additionally interesting to examine what happens with the gesture of pitch-contol in short voiceless parts, because often local extreme values in F0 can be found at the borders of voiced and unvoiced parts. But often these are parts where the pitch determination algorithms do not work very reliable. However, we assume that a longer melodic struture, gesture or intention is broken sometimes by unvoiced phonemes. So we suggest to define a paramater called *virtual F0*. This parameter could be found by appropriate methods of interpolation in the voiceless parts like cubical splines etc. Furthermore, this an assumption can help us to improve the reliability of the statistics, and simplifies the prediction of F0 curves.

In the next step F0 contours were performed by manual straight-line stylization corresponding to the mentioned assumptions that typical F0 values, intervals and accents were speaker-typical. Exploring and collecting local extreme values by closed-copy stylization of F0 does indeed results in significant speaker-dependent intervals and values occuring often for a falling after accents. Examining all local extreme values including the virtual accents, a set of typical values of maximum F0 (H*-tone) and minimum F0 (L-tone) for each speaker can be found. Some of them are describing typical ranges and intervals for each speaker.

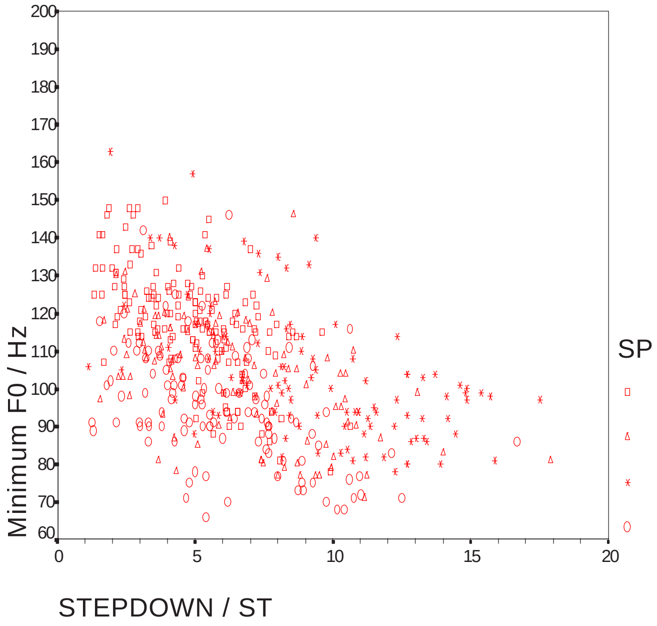

Fig. 4: min. F0 and falling intervals for each speaker (SP)

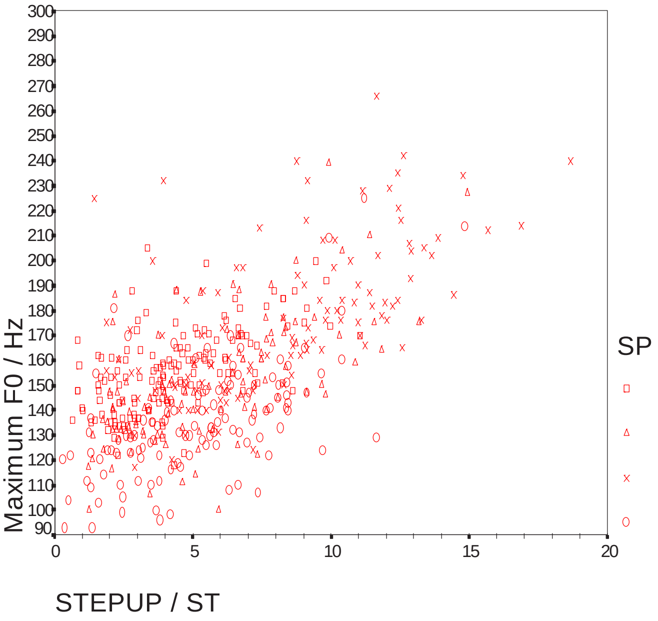

Fig. 5: max. F0 and rising intervals for each speaker (SP)

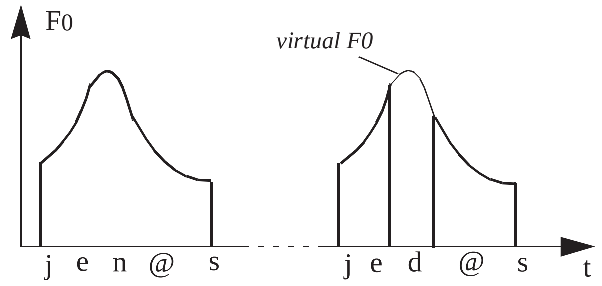

Fig. 6: An Example for a *virtual F0 (jenes.., jedes)*

5. DISCUSSION

The mentioned features can be extracted and transplanted successfully to other materials, using an *Intonator* [4]. It provides a very convenient interactive graphic tool for painting F0 contours and resynthesis by PSOLA-technique or intonation-melody. It is influenced by the IPO-method already developed the late sixties [8]. A first perceptual evaluation pointed out that the approximated and resynthesized F0 contours sound very natural and are representative for the speakers concerned. The declination, mean and variance of F0 can be treated as a resulting effect of these characteristical tones. Exploring the contours visually a weak decrease in variance during the text and sentences can be found, because the speakers were non-professional. Perceptual strategies of subjects and training effects should be reviewed and analysed in more detail. In principle the model should be language-independent. Superposition by other effects, like context and speaking styles, cannot be excluded generally. In future the model will be linked to a TTS system. Sometimes the PSOLA algorithm causes problems, because large intonational manipulation results in obvious spectral changes.

6. CONCLUSION

The perceptual results of transplanting the presented parameters to other speakers are very promising. An automatic analysis and generation modul is in preparation. For further listening experiments information is now available regarding the relevant terms of human intonational speaker recognition. The experimental processing can be extended to other prosodical analyses and to the evaluation of e.g. speaking-styles, emotion, etc. If the presented results hold true for other speech material, the next step will be to integrate the model into the prosody generation part of a TTS system. A similar process, directed towards the speaker-dependent perception of rhythm and segmental duration, is in preparation.

ACKNOWLEDGEMENTS

The presented studies are part of a research project in speaker-modelling and were supported by DFG (German Federal Research Found).

REFERENCES

[1] Collier R. (1996), “*Prosodic analysis: A Dual Track?*“, in: J.P.H van Santen et al. (ed.) Progress in Speech Synthesis, Springer New York 1997

[2] Larkey L.S. (1983), “*Reiterant speech: An acoustic and perceptual validation*“ J. Acoust. Soc. Am. 73(4), pp. 1337-1345

[3] Nicholas P., Romeas P. (1993), “*Evaluation of Prosody in the French Version of Multilingual Text-to-Speech synthesis: Neutralising Segmental Information in Preliminary Tests*“, Int. Proc. EUROSPEECH’93 Berlin 1993, pp. 211-214.

[4] Mariniak A., Mersdorf J. (1994), „*Ein Ansatz zur Beurteilung der Intonation von synthetischer Sprache*", in: Fortschritte der Akustik - DAGA ’94, DPG Verlag Bad Honnef, 1369-1372.

[5] Sonntag, G., Portele, T. (1996), “*A framework for evaluating and verifying the presence of linguistic concepts in the prosody of spoken utterances*“, Proc. 3 of SPEAK!-Workshop, Budapest 1996 (to appear)

[6] Mersdorf J. „*Ein Hörversuch zur perzeptiven Unterscheidbarkeit von Sprechern bei ausschließlich intonatorischer Information*", in: Fortschritte der Akustik - DAGA ’96, DPG Verlag Bad Honnef, 482-483.

[7] Higuchi N., Hirai T., Sagsaka Y. (1996), “*Effect of Speaking Style on parameters of fundamental Frequency Contour*“, in: J.P.H van Santen et al. (ed.) Progress in Speech Synthesis, Springer New York 1997

[8] ’t Hart J., Collier R., Cohen A. (1990), “*A Perceptual Study of Intonation*“, Cambridge Studies in Speech Science and Communication, Cambridge University Press.

[9] Sendlmeier W.F., Holtzmann U. (1991) „*Sprachqütebeurteilung mit Passagen fliessender Rede*", in: Fortschritte der Akustik - DAGA ’91, DPG Verlag Bad Honnef, 969-972.

[10] Vincent M., DiChrito A., Hirst D. (1995), “*Prosodic Features of Finality for Intonation Units in French Discourse*“ in: Procceddings ICPhS Stockholm, 1995 pp 2.718-721.