Keynote Speaker 2

Five Dimensions of Prosody: Intensity, Intonation, Timing, Voice Quality, and Degree of Reduction

Hartmut R. Pfitzinger, University of Munich

Wednesday, May 3, 09:00 – 09:45

Extended Abstract

This talk gives an overview of methods for analysis, modification, and synthesis of the prosodic properties of speech. The term prosodic properties is supposed to cover all phenomena that are not segmental and that are described on several tiers parallel to the segmental tier. Firth [1] called them prosodies and according to him e. g. distant assimilation (such as Turkish vowel harmony) is also covered by his term. This very broad meaning of the term prosodies is much closer to my view than the very common habit of saying prosody and meaning only intonation. One of the main purposes of this talk is to demonstrate that the manipulation of intonation or timing alone can sometimes produce prosodically contradictory stimuli which in turn inconsistently degrade perception results.

I entitled the talk "Five Dimensions of Prosody" because the first three dimensions intensity, intonation, and timing are very well known, and Campbell and Mokhtari [2] named voice quality the fourth prosodic dimension. Obviously, I have added another dimension, one which I would like to name the degree of reduction. The remainder of the talk is concerned with describing the analysis, modification, and synthesis of each of these five prosodies.

Intensity be measured easily reliably by means of root–mean–square, or rectifying and averaging, or, more precisely, by smoothing the instantaneous amplitude achieved via Hilbert transformation. It is often argued that intensity, whether short–term or long–term, has obviously minor communicative functions as can be seen from any broadcasting where the short–term amplitude is generally strongly manipulated to maximize loudness of speech without any noticeable impact on the meaning of the speech. But intensity should be taken into account when naturalness is important (e. g. in speech synthesis) or e. g. when perception stimuli with shifted word accents are necessary. In this case the shift of the intonation peak should be accompanied by a shift of an intensity peak.

Intonation is more complicated: although there exist countless F0 or glottal epoch detection algorithms (Hess [3] gives an overview), none of them is absolutely reliable and many of them work only on high–quality speech recordings, or on a limited range of F0 values, or use an inferior voiced/unvoiced–detection method, or are sensitive to amplitude variations, or suffer from other shortcomings. However, it turned out that error rates of modern algorithms are sufficiently low and many times cancelled by post–processing. Subsequent smoothing, extrapolation, and parameterization are necessary to make intonation accessible to meaningful modification, each of these methods with its own algorithmic problems. It turned out

that the command–response model of Fujisaki allows for a very powerful parameterization of intonation contours in many languages [4]. Finally, pitch–synchronous overlap and add (PSOLA) is a very effective way to synthesize the new signal but it has problems especially with strong F0–changes and high–pitched female voices.

Timing is even more complicated: In 1998 [5] I invented a model to estimate perceptual local speech rate (PLSR), a prosodic contour similar to F0 contours, easy to interpret and to modify. It is based on a linear combination of the local syllable rate and the local phone rate both of which are estimated from manual segmentations of phones and syllable centers, and it produces a mean deviation of 10 percent of the perceptual speech rate which is precise enough for phonetic studies and speech synthesis. Other methods such as Z–score–based duration contours [6] suffer from probable inconsistencies in their essential knowledge bases, i. e. prototypical mean durations and standard deviations of all speech segments, and from nonlinear elasticity of the segments. For simple copy–synthesis, dynamic time warping (DTW) is more appropriate since it needs no segmentation of both speech signals. Speech pauses, and especially hesitations and repairs further complicate the timing structure of speech. Their positions and durations are difficult to predict [7].

Voice quality is difficult to measure, modify, and synthesize. Convincing approaches are based on epoch detection methods and inverse filtering techniques, two significant sources of error. The goal is to obtain and parameterize the glottal–flow waveform which is supposed to carry all voice quality properties. While the paper of Campbell and Mokhtari [2] is based only on the normalized amplitude quotient (NAQ), which represents a continuum from breathy to modal or even pressed voice quality, a more holistic approach of Mokhtari, Pfitzinger, and Ishi [8] consisted in applying a principal components analysis (PCA) to a database of glottal–flow waveforms for the purpose of later reconstructing and interpolating all underlying glottal–flow waveforms from just a few principal components (PCs). A starting point to cover a wide range of laryngeal variations was the typology of phonation by Laver [9] and his recordings. It turned out that the first PC mainly accounted for F0 variations which raises the question as to whether the prosodic dimension intonation is better subsumed under the fourth dimension since variations of F0 also influence voice quality.

The degree of reduction is hardly ever interpreted as a prosody, and for good reason: in order to estimate the degree of reduction a huge amount of phonological knowledge is necessary. That is, the canonical form must be known for any utterance to count the number of elisions and insertions (effects in the time domain), and the target formant frequencies (or articulatory target positions) must be known to estimate the segmental undershoot (frequency domain effect). This is closely related to Lindblom's HH theory of phonetic variation [10]. One problem is that there is not only purely mechanical coarticulation, constrained by inherent properties of the speech organs, but also coarticulation rules learned during language acquisition. Even though this prosody is very difficult to estimate, first approaches are highly desirable since it is very important when manipulating speech. E. g. shifting the word accent from one syllable to another is a real problem because the former unstressed syllable usually is produced in a strongly reduced way (in English often as a Schwa) while the stressed syllable generally has a non–central vowel quality and a longer duration. Thus, the target syllable should become

de–reduced and the source syllable reduced.

It should be clear that each of the above–mentioned prosodies has its segmental and supra–segmental manifestation. Actually, from my point of view the terms low–frequency components and high–frequency components of prosodies describe the speech facts in a better way. In this view, even the articulatory movements, and thus every detail that constitutes speech, could become prosodies.

At the end of the talk two applications of prosodic modifications are demonstrated: one is speech morphing between two utterances of different speakers, which means estimating equally–spaced intermediate utterances with all prosodic properties changing in equal steps from one speaker to another. And the other application is in the field of computer–aided language learning (CALL). Here, we try to show that an automatic prosodic correction of the speech signal of a language learner and its auditory feedback help the learner to aquire a foreign language faster than by hearing the corrections spoken with the teacher's voice [11].

References

[1] Firth, J. R. (1948). Sounds and prosodies. Transactions of the Philological Society, pp. 127-152.

[2] Campbell, N.; Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In Proc. of the XVth Int. Congress of Phonetic Sciences, vol. 3, pp. 2417-2420, Barcelona.

[3] Hess, W. (1983). Pitch determination of speech signals: Algorithms and devices. Springer-Verlag, Berlin, Heidelberg, New York.

[4] Fujisaki, H. (2004). Information, prosody, and modeling with emphasis on tonal features of speech. In Proc. of the 2nd Int. Conf. on Speech Prosody, pp. 1-10, Nara; Japan.

[5] Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In Proc. of ICSLP '98, vol. 3, pp. 1087-1090, Sydney.

[6] Campbell, W. N. (2000). Timing in speech: A multi-level process. In Horne, M., ed., Prosody: Theory and experiment. Studies presented to Gösta Bruce, pp. 281-334. Kluwer Academic Publishers, Dordrecht.

[7] Pfitzinger, H. R.; Reichel, U. D. (2006). Text-based and signal-based prediction of break indices and pause durations. In Proc. of the 3rd Int. Conf. on Speech Prosody, Dresden; Germany.

[8] Mokhtari, P.; Pfitzinger, H. R.; Ishi, C. T. (2003). Principal components of glottal waveforms: Towards parameterisation and manipulation of laryngeal voice-quality. In Proc. of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (Voqual'03), pp. 133-138, Geneva.

[9] Laver, J. (1980). The phonetic description of voice quality. Cambridge University Press, Cambridge.

[10] Lindblom, B. E. F. (1990). Explaining phonetic variation: A sketch of the HH theory. In Hardcastle, W. J.; Marchal, A., eds., Speech production and speech modelling, Nr. 55 in Nato ASI series D: Behavioural and social sciences, pp. 403-439. Kluwer Academic Publishers, Dordrecht, Boston, London.

[11] Bissiri, M. P.; Pfitzinger, H. R.; Tillmann, H. G. (2006). Lexical stress training of German compounds through resynthesis and emphasis. Accepted for Proc. of InSTIL Workshop at CALICO, Hawaii.

About the Speaker

Hartmut R. Pfitzinger was born in September 1968 in Kiel, Germany. From 1989 to 1992 he studied Phonetics and Digital Speech Processing (Prof. K. Kohler), Computer Science, Mathematics, and Psychology at Christian-Albrechts-Universitaet in Kiel. From 1992 to 1995 he studied Phonetics and Speech Communication (Prof. H. G. Tillmann), Computer Science, and Computational Linguistics at Ludwig-Maximilians-Universitaet Munich.

In 1995 he obtained the M.A. degree (title of thesis: "Vowel Perception and Acoustic Features") and started working as a scientist and lecturer at the Institute of Phonetics and Speech Communication, Munich. His duties comprised the consulting or leading of several spoken language collection projects which were conducted in cooperation with the companies Siemens AG, Philips, and Volkswagen AG. Recently, he was head of a basic research project on speech synthesis supported by the BMW Group.

In 2001 he received the Ph.D. degree (title of thesis: "Phonetic Analysis of Speech Rate"). From February to May 2003 and 2004 he worked as an invited scientist in the Japan Science and Technology Corporation's (JST) project entitled Expressive Speech Processing (ESP) with the leader of the project Dr. N. Campbell at the Advanced Telecommunications Research Institute International (ATR). During this time he was concerned with the parameterization of the excitation signal and with speech-to-speech synthesis. His main research topics are perception, modification, and synthesis of speech.