Towards a Model of Target Oriented Production of Prosody

Grzegorz Dogil, Bernd Möbius

Institute of Natural Language Processing—Experimental Phonetics
University of Stuttgart, Germany
{dogil,moebius}@ims.uni-stuttgart.de

Abstract

A new paradigm for prosody research is presented, inspired by the speech production model recently proposed by Guenther, Perkell, and colleagues. This research paradigm aims at generalizing the production model by extending it from a predominantly segmental perspective to a new theory of the production of prosody. Speech movements in the prosodic domain are interpreted as intonational gestures that are planned to reach and traverse perceptual target regions. Evidence from F_0 alignment studies suggests that the perceptual targets can be approximately represented by regions in a multidimensional acoustic-temporal space. These studies also indicate that segmental, spectral, temporal, and prosodic structure are co-produced in such a way as to mutually support and enhance, and not impair, the perceptual targets. Furthermore, examples of multilevel mappings between invariant and variable targets in the domain of prosody are provided, and a dichotomy of phonemic and postural prosodic settings is discussed.

1. Introduction

Prosody has an integrating function in the organization and production of speech, by embedding semantic information (intonational meaning), syntactic structure (phrasing), morphological structure (metrical spellout), and segmental sequences (segmental spellout) into a consistent set of address frames (syllables, metrical feet, phonological words, intonational phrases) [1, 2]. However, this important role of prosody is not usually reflected in speech production models.

In this paper a novel avenue of research into the production of prosody is proposed. This new research paradigm uses Guenther's and Perkell's speech production model [3, 4, 5] as a point of departure, moving from the predominantly segmental perspective to a new theory of the production of prosody. The model underlying the proposed approach posits that speech production is constrained by auditory and perceptual requirements. The only invariant targets of the speech production process are auditory perceptual targets.

In our extension and generalization of the model, speech movements in the prosodic domain are interpreted as intonational gestures that are planned to reach and traverse perceptual target regions. The targets are characterized as multidimensional regions in the perceptual space. Gestures that are successfully executed by the speaker produce acoustic realizations of perceptually relevant prosodic events, such as those predicted by intonational phonology. Examples of mapping relations between reference frames (the target regions) and intonational gestures are discussed in section 2.

The prosodic interpretation of the speech production model is structured around a hierarchy of prosodic domains, the core of which is also a conventional foundation for the linguistically based classification of intonational events. The top level of this hierachy is *discourse structure*, phonologically represented and phonetically realized by pitch range and register. Discourse segments have an internal *information structure*, represented by topic and focus, which signal given/new relations. Utterances and phrases are assumed to have an *accentual structure*, phonologically represented as tones and phonetically realized as pitch accents and boundary or phrase tones. The phonological elements (the tones) are systematically combinable into tunes, which may carry meaning and can thus be attributed a morphological status.

Orthogonal to this hierarchy, and pervading it, we again follow Guenther [4] by positing a dichotomy of *phonemic settings* and *postural settings* (section 3). In mature speech production auditory feedback has two functions. First, it helps maintain phonemic settings, i.e. parameters of phonemic distinctions; second, it assures intelligibility by monitoring the acoustic environment and accommodating the baseline postural settings of the respiratory, laryngeal, and supraglottal systems appropriately.

We further advocate the implementation of computational prosody models that have their motivation both in the theory of speech production and in linguistic theory. These computational models are intended to serve two main purposes. First, they will allow us to empirically test the assumptions made by the production model, for instance the effect of speaking rate and other factors related to speech timing on the acoustic realization of intonational gestures. Second, the linguistically based classification of intonational events, e.g. those related to (a) discourse structure (register, pitch range), (b) information structure (topic, focus), and (c) accentual patterns (pitch accents, tones, tunes), can be experimentally tested by trainable intonation event classifiers.

2. Mapping relations

In his *blueprint of the speaker*, Levelt [6] has framed a research program whose agenda calls for the design and implementation of empirically viable working models of the processing components involved in the blueprint. We suggest that computational models (e.g., neural nets) can learn the mappings between reference frames (the perceptual target regions) and intonational gestures in speech [4, 7], i.e., they can learn the optimal settings of phonetic details and articulatory gestures that are related to specific intonational events.

We expect such learned computational mappings, based on adaptive weights, to tend towards consistent realization configurations, either under the influence of temporal constraints or as part of a tune (a coherent sequence of intonational events), as long as the perceptual target region is traversed. The postural relaxation hypothesis [4] posits that in order to achieve a

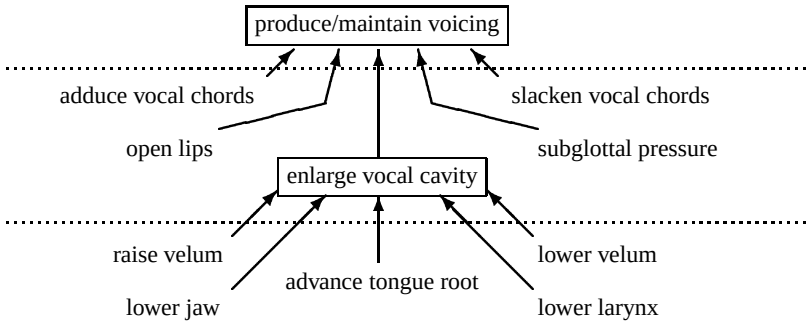

Figure 1: *An invariant acoustic-perceptual goal (here: maintaining voicing) may be achieved by multi-level variable articulatory goals.*

perceptual goal, the speaker is consistently biased toward certain articulatory gestures. The computational simulation plays a crucial role in providing evidence for the postulated internal models involved in speech production.

To illustrate our methodological approach in the domain of prosody, let us consider two examples: modeling the production and maintenance of voicing (section 2.1), and modeling pitch accents (section 2.2).

2.1. Voicing

One basic prerequisite for intonational gestures is the production of F_0 contours by the voice source. The invariant acoustic and perceptual target is thus the presence of voicing, which can be characterized in the acoustic domain as a quasi-periodic structure containing a considerable amount of low-frequency energy. In the articulatory domain voicing is produced by an approximately periodic vibration of the vocal chords.

Vocal chord vibration is a fragile and complex process; it tends to break down unless specific conditions are satisfied. Figure 1 illustrates how the invariant acoustic-prosodic target can be mapped onto a number of variable articulatory goals, which may be characterized by the following gestures, all of which are known to support voicing: (a) adducing the vocal chords; (b) slackening the vocal chords; (c) generating sufficient subglottal pressure; (d) producing a sufficient degree of mouth opening; (e) enlarging the vocal cavity. Each of these articulatory goals contributes to the maintenance of voicing, but in practice voicing is maintained by a combination of most or all of these gestures, and there exist several trading relations between the individual gestures [8].

The articulatory goal of enlarging the vocal cavity can in turn be achieved by a number of second-level variable articulatory goals, which may be described as follows: (e1) raising the velum; (e2) lowering the velum; (e3) advancing the tongue root; (e4) lowering the jaw; (e5) lowering the larynx. Again, trading relations exist between several of these goals, and some of them are in fact mutually exclusive, e.g. (e1) and (e2).

We suggest that a computer simulation in the form of a neural network architecture may learn the optimal mappings from the invariant acoustic and perceptual targets to the level-one and level-two variable articulatory targets.

2.2. Accents

Another example for a multi-level mapping between invariant and variable targets is illustrated in Figure 2. On the acoustic level, eight distinct F_0 contours are shown that a statistical classifier extracted from a large training database and entered as prototypes into a codebook [9].

The optimal mapping of these acoustic prototypes onto a small number of phonologically, and thus perceptually, motivated intonational categories (pitch accents or tones, here: L*H oder H*L) may be acquired by means of a supervised learning procedure. Furthermore, each codebook entry is characterized by a set of parameters that determine its concrete shape in the time and frequency dimensions. These parameters can in turn be mapped onto a number of articulatory gestures. F_0 can be raised by increasing the subglottal air pressure, by raising the back of the tongue, by increasing vocal chord tension, or by activating certain intrinsic laryngeal muscles; lowering F_0 is achieved by opposite gestures. The speaker may use these individual gestures, as well as combinations of gestures, to produce rising or falling F_0 movements.

Again, we are dealing with a multi-level mapping of invariant perceptual targets on variable goals. This time the variable goals are located partly on the acoustic and partly on the articulatory level of manifestation.

2.3. Mutual enhancement

In the prosodic domain the intended perceptual impression is produced by the precisely orchestrated mutual enhancement of several acoustic parameters, such as duration and spectral energy and its distribution within the pertinent prosodically relevant category (e.g., syllable or accent group). Recent studies of the temporal alignment of F_0 contours have shown that speakers carefully control the synchronization of F_0 with the segmental structure, taking into account the segmental composition of syllables and accent groups as well as the durations of the pertinent segments and syllables [10].

It has also been demonstrated that the optimal perception of prosodic contrasts takes place at those locations in the speech signal where other acoustic parameters, such as formant frequencies, bandwidths and amplitudes, remain relatively stable [11]. Moreover, a falling F_0 movement is perceived as either a low, falling, or high tone, depending on its temporal alignment with the segmental structure of the syllable that it is associated with and the amount of new spectral information that this segmental structure introduces at a given point in time [12].

An F_0 movement that characterizes a particular accent is thus properly perceived only if, firstly, the duration of the category carrying the accent is optimally adjusted and, secondly, no new spectral information is introduced by the segmental struc-

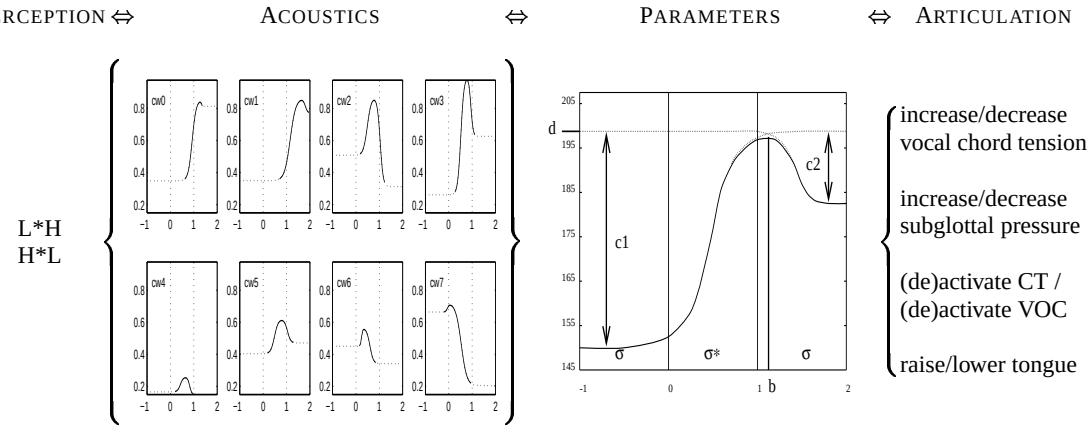

Figure 2: *Multi-level mapping between invariant perceptual targets (pitch accents or tones, here: L*H or H*L) and variable acoustic (F_0 prototypes and their parameterizations) and articulatory goals.*

ture at the crucial location. In order to convey the prosodically relevant contrasts, the segmental acoustic parameters need to be distributed in the temporal domain in such a way as to support and enhance, and not impair, the perceptual targets of prosody, for instance accent-lending F_0 movements and spectral tilt changes.

3. Phonemic and postural settings

Phonemic settings, once learned, tend to be stable and resistant to change. Evidence comes from the investigation of intonational foreign accent, which has been shown to be partly rooted in the stable phonological representation of prosody of the first language [13], as well as from studies that show that a speaker's vowel space remains stable after adult hearing loss [5]. Certain prosodic gestures are also more resistant after hearing loss than others, and we hypothesize that the resistant ones are those whose function is to make phonological distinctions. For instance, speakers with adult hearing loss continue to use stress linguistically: the learned internal model of stress, along with the articulatory gestures and resulting acoustic correlates of stress, remains stable.

Postural settings, in contrast, are lost much earlier. In general, problems related to suprasegmental properties of speech, such as intensity (sound pressure level) and F_0 control, and speaking rate, are usually observed soon after hearing loss [14]. Experiments with manipulated F_0 feedback point in the same direction [14]. F_0 control partly relies on closed-loop feedback to achieve a pitch target [15].

It thus appears that these findings partially contradict each other: for instance, F_0 is observed to be both stable, as manifested by intonational foreign accent, and unstable, e.g. after hearing loss. We suggest that this apparent contradiction can be resolved by analytically separating two properties of prosodic parameters. The first property pertains to phonemic settings: it involves the linguistically relevant and phonologically distinctive functions of prosodic features, e.g. accent as a focus marker. The second property pertains to postural settings, i.e. to the role that the prosodic parameters play in the continuous adjustment of overt speech, based on closed-loop auditory feedback. The postural parameters can be changed rapidly by speakers with normal hearing to adapt to varying acoustic conditions; this adaptation capability is lost soon after hearing loss. The learned internal model of phonemic settings does not rely on continuous auditory feedback and parameter update and is thus far more robust.

The intended analytical separation is expected to be difficult because of *interactions* between postural changes and phonemic settings. For instance, F_0 is generally controlled through moment-to-moment feedback and with reference to an internal pitch representation [14]. Text coherence (discourse and utterance intonation) is known to be lost early but intra-syllabic settings (tones, pitch accents) tend to be stable, even though the parameter F_0 is involved in both domains.

We suggest that quantal effects are involved in achieving the relative stability of phonemic settings. There is a region in which a given parameter of overt speech is stable even if speech production gestures and movements are exectued with some inaccuracy, an observation that is consistent with the theory of the quantal nature of speech [16, 17].

Stability characterizes most speech sounds, and phonemic instability is the exception. Moreover, it has been hypothesized that invariant phonetic shapes are protected by sound laws and less likely to undergo sound change [18]. We posit that this property of speech pertains not only to the segmental domain but to the prosodic domain as well.

For instance, the stability of the F_0 output that is correlated with the realization of tones is enhanced by aligning the F_0 target with an area of minimal spectral change [11]. This requirement needs to be balanced with a conflicting constraint, viz. that tones be aligned in relative vicinity to "pivots". The pivot is an area at which the maximum of new spectral information coincides with rapidly rising intensity [19], e.g. in consonant-vowel transitions. The new information causes an onset of auditory firing, and gestures realized in the vicinity of this onset are perceptually more salient than in other areas.

4. Conclusion

A new paradigm for prosody research has been presented that is inspired by the speech production model recently proposed by Guenther and Perkell [3, 4, 5]. This research paradigm aims at

generalizing the production model by extending it from a predominantly segmental perspective to a new theory of the production of prosody.

In the prosodic domain we interpret speech movements as intonational gestures that are planned to reach and traverse perceptual target regions. The targets may be characterized as multidimensional regions in the perceptual space. Gestures that are successfully executed by the speaker produce acoustic realizations of perceptually relevant prosodic events. While we do not intend to identify perceptual with acoustic goals, we assume that perceptual targets can be approximately represented by target regions in a multidimensional acoustic-temporal space. However, the relation between invariant perceptual targets and acoustic properties of the speech signal remains an important research topic.

Mapping relations between reference frames (the target regions) and intonational gestures and the methodological approach were illustrated by the two cases of, first, modeling the production and maintenance of voicing and, second, modeling pitch accents. Both examples involved a multi-level mapping between invariant and variable targets in the domain of prosody.

It was further argued that two types of prosodic parameters need to be distinguished, the first type pertaining to phonemic settings, which involve internal models of phonologically distinctive functions of prosodic features, and the second type pertaining to postural settings, i.e., to the role that the prosodic parameters play in the continuous adjustment of overt speech, based on closed-loop auditory feedback. Given the evidence for such a dichotomy on the segmental level, we suggest that the mechanisms involved in the acquisition and control of prosody may not differ categorically from those that control segmental speech production.

The important integrating role of prosody is not adequately reflected in most speech production models, and the new research paradigm is intended to overcome this shortcoming. We also believe that the proposed model has the potential to bridge the gap between the currently prevailing phonologically motivated intonation theories, i.e. autosegmental-metrical (or tone-sequential) models of intonation, and the production oriented model of intonation advocated by Fujisaki [20, 21].

5. References

[1] Levelt, W. J. M., *Speaking: From Intention to Articulation*, MIT Press, Cambridge, MA, 1989.

[2] Dogil, G., "Understanding prosody," in *Psycholinguistics—An International Handbook*, Rickheit, G., Hermann, T., and Deutsch, W. (Eds.), de Gruyter, Berlin, 2000.

[3] Guenther, F. H., "A modeling framework for speech motor development and kinematic articulator control," in *Proc. 13th Internat. Congress of Phonetic Sciences (Stockholm)*, 1995, 2:92–99.

[4] Guenther, F. H., Hampson, M., and Johnson, D., "A theoretical investigation of reference frames for the planning of speech movements," *Psychological Review*, 105:611–633, 1998.

[5] Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., and Zandipour, M., "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," *Journal of Phonetics*, 28 (3):233–272, 2000.

[6] Levelt, W. J. M., "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, Brown, C. M., and Hagoort, P. (Eds.), 83–122. Oxford University Press, Oxford, UK, 1999.

[7] Möhler, G., and Conkie, A., "Parametric modeling of intonation using vector quantization," in *Proc. Third ESCA Workshop on Speech Synthesis (Jenolan Caves, Australia)*. 1998, 311–316.

[8] Jessen, M., "Phonetic implementation of the distinctive auditory features [voice] and [tense]," *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, 6(4):11–62, 2000.

[9] Möhler, G., Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese, Ph.D. thesis, *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, 4(2), 1998.

[10] van Santen, J. P. H., and Möbius, B., "A quantitative model of F_0 generation and alignment," in *Intonation—Analysis, Modelling and Technology*, Botinis, A. (Ed.), 269–288. Kluwer, Dordrecht, 2000.

[11] House, D., *Tonal Perception in Speech*, Lund University Press, Lund, 1990.

[12] House, D., "Differential perception of tonal contours through the syllable," in *Proc. Internat. Conf. on Spoken Language Processing (Philadelphia, PA)*, 1996, 1:2048–2051.

[13] Jilka, M., The contribution of intonation to the perception of foreign accent, Ph.D. thesis, *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, 6(3), 2000.

[14] Jones, J. A. and Munhall, K. G., "Perceptual calibration of f0 production: evidence from feedback perturbation," *Journal of the Acoustical Society of America*, 108(3):1246–1251, 2000.

[15] Titze, I., *Principles of Voice Production*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[16] Stevens, K. N., "The quantal nature of speech: evidence from articulatory-acoustic data," in *Human Communication: A Unified View*, Davis, E. E., and Denes, P. B. (Eds.), 51–66. McGraw-Hill, New York, 1972.

[17] Stevens, K. N., "On the quantal nature of speech," *Journal of Phonetics*, 17:3–45, 1989.

[18] Dogil, G., and Möhler, G., "Phonetic invariance and phonological stability: Lithuanian pitch accents," in *Proc. Internat. Conf. on Spoken Language Processing (Sydney)*, 1998, 7:2891–2894.

[19] Dogil, G., "Acoustic landmarks and prosodic asymmetries," in *Proc. 14th Internat. Congress of Phonetic Sciences (San Francisco)*, 1999, 3:2105–2108.

[20] Fujisaki, H., "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The Production of Speech*, MacNeilage, P. F. (Ed.), 39–55. Springer, New York, 1983.

[21] Fujisaki, H., "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal Physiology: Voice Production, Mechanisms and Functions*, Fujimura, O. (Ed.), 347–355. Raven, New York, 1988.