Phonemic and Postural Effects on the Production of Prosody
Bernd Möbius & Grzegorz Dogil
Institute of Natural Language Processing
University of Stuttgart, Germany
{Bernd.Moebius,Grzegorz.Dogil}@IMS.Uni-Stuttgart.DE

Abstract
Phonemic settings and the internal models that they represent are learned in the process of language and speech acquisition. Postural settings, in contrast, rely on continuous auditory monitoring and tend to break down quickly if this monitoring process is inhibited during speech production. Evidence presented in the literature seems to indicate that stable internal models are mostly associated with segmental phonemic targets, whereas prosodic features often display postural characteristics. In this paper it is argued that the dichotomy of phonemic and postural settings applies not only to segmental properties of speech but to prosodic features as well. Phonemic and postural effects on the production of prosody are reviewed and it is suggested that the boundary between phonemic and postural effects on a given prosodic feature is flexible. We further hypothesize that the speaker may rely on a set of acquired internal models and select from this set a particular model depending on communicative and situative constraints.

1. Introduction
This work is part of a new research paradigm that builds upon the speech production model by Guenther, Perkell, and colleagues [1, 2, 3]. This model posits that speech production is constrained by auditory and perceptual requirements. The only invariant targets of the speech production process are assumed to be regions in auditory perceptual space.

We have proposed an extension and generalization of Guenther’s and Perkell’s model, by integrating its segmental perspective with a new theory of the production of prosody [4]. According to this new approach, speech movements in the prosodic domain are interpreted as tonal and temporal 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 tonal gestures were also presented [4].

Our prosodic interpretation of the speech production model is structured around a hierarchy of prosodic domains, comprising discourse structure, information structure, and accentual structure. Orthogonal to this hierarchy—we follow Perkell and Guenther [3] again—a dichotomy of phonemic settings and postural settings is posited. 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.

There is now an increasing body of evidence (see [5] for an overview) that auditory feedback plays an important role in the implementation, programming, planning, and monitoring of prosody in speech production. It is therefore tempting to associate phonemic settings and the internal models that they represent with the segmental domain, and to attribute mostly postural effects to parameters in the prosodic domain, such as fundamental frequency ($F_0$), speaking volume, and speaking rate.

It is our aim in this paper to argue that the postural/phonemic dichotomy applies equally to segmental and prosodic properties of speech and that it pervades the above-mentioned hierarchy of prosodic domains. We further hypothesize that there may be more than one type of internal model for the speaker to rely on and that the speaker may select one model out of a set of acquired internal models depending on communicative and situative constraints.

2. Phonemic and postural settings
Perkell and colleagues suggest that the role of auditory feedback in the planning of speech movements is different for phonemic and postural settings [3].

Auditory feedback is used to acquire and later update the mapping from articulatory gestures to auditory trajectories. It does not serve to continuously monitor the current state of the vocal tract with respect to the auditory space because such a monitoring would introduce delays that are prohibitive during actual speech production.

According to the speech production model there is a unique phonetic target region in auditory-temporal space for each phoneme of a given language [6]. The DIVA model [1, 7] provides two processing mechanisms that perform unidirectional mappings (one forward, one inverse) between abstract phonemes and the auditory targets that represent them. These mappings are necessarily both language specific and phoneme specific.

The process of language and speech acquisition involves establishing the phonemic mappings. Once learned, phonemic settings tend to be stable and resistant to change. This is evidenced by studies showing that a speaker’s vowel space remains stable after adult hearing loss. Even in the absence of hearing the vowel space, as well as consonantal phonemic distinctions such as the contrast between /s/ and /ʃ/, remain largely congruent with normative patterns even years after onset of hearing loss [3].

In the prosodic domain, the stability of phonemic settings is evidenced by results of studies on intonational foreign accent, which has been shown to be partly rooted in the stable phonological representation of prosody of the first language [8]. Furthermore, certain prosodic gestures are 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.

Whereas phonemic settings are assumed to be stable even after a change in hearing status, such as adult hearing loss or acquisition of hearing in conjunction with a cochlear implant, postural settings are apparently prone to become destabilized more easily and significantly faster. 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 [3, 9]. Experiments with manipulated \( F_0 \) feedback point in the same direction [9]. \( F_0 \) control partly relies on closed-loop feedback to achieve a pitch target [10].

It would be premature, however, to conclude that prosodic features of speech generally belong into the realm of postural settings. The role of stable internal models of prosody as a cause of intonational foreign accent indicates that language learners acquire internal prosodic models along with segmental ones. Recent follow-up experiments involving modified \( F_0 \) feedback further suggest a more complicated interpretation of the control of prosodic speech parameters (see section 3.1).

The findings concerning \( F_0 \) control appear to contradict each other partially. \( F_0 \) is reported to be both stable, as manifested by intonational foreign accent, and unstable, for instance 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, such as accent as a focus marker. The second property pertains to postural settings, 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 [9]. Text coherence (discourse and utterance intonation) is known to be lost early, but local prosodic settings (tones, pitch accents) tend to be stable, even though the parameter \( F_0 \) is involved in both domains.

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 [11]. We posit that this property of speech pertains to the prosodic domain.

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 [12]. 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 [13], such as 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.

3. Effects on prosodic features

In this section we discuss the effects on prosodic features exerted by several independent factors, such as speaking style and the acoustic environment. The prosodic features under consideration comprise \( F_0 \) as the acoustic correlate of tonal features; sound pressure level (SPL) as an acoustic correlate of amplitude features; and speaking rate, phrase breaks and segmental durations as correlates of temporal features.

3.1. Tonal features

The production of \( F_0 \) is monitored and controlled through a closed-loop negative feedback system: when subjects are exposed to artificially raised or lowered pitch of auditory feedback, they compensate for the difference between intended and perceived pitch by changing \( F_0 \) in the opposite direction [9, 14]. There is a latency (>150 ms) of the compensating response corresponding to the time consumed by the tasks of processing the auditory feedback signal, modifying efferent laryngeal control, and changing settings of the laryngeal structures. Because of the latency the compensation exerts its main effect on word, phrase, and utterance prosody rather than on the syllable in which the change of pitch feedback is induced [14].

Evidently, there is an upper bound on the magnitude of the frequency shift: responses seem to be limited to a maximum of 60 cents (1 cent = 1/100 semitone). The existence of an upper bound is explained by the integration of two feedback channels, viz. auditory and proprioceptive. In the pertinent experiments the proprioceptive channel signals an appropriate laryngeal setting for the actual \( F_0 \) production, whereas the auditory feedback signals a deviation. Discrepancies between the two feedback channels may introduce a nonlinear threshold on the compensation effort [14].

If the shifted \( F_0 \) feedback experiments are carried out with trained singers who are singing scales without any external reference, the unpredictably introduced \( F_0 \) feedback manipulation is completely compensated for [15]. Notice that trained singers do not have to rely on an external reference; they may use reference frequencies that are represented internally. Our interpretation of this finding is that trained singers acquire internal tonal models (scales) that are as stable and resistant to disturbances as are learned phonemic models in speech production.

The shifted \( F_0 \) feedback experiments discussed above were carried out with speakers of English, where syllable and word level pitch is viewed as a predominantly postural cue which does not primarily subserve phonological functions.\(^1\) Tone languages, on the other hand, use contrastive \( F_0 \) patterns on the syllable level to mark lexical tone; \( F_0 \) thus serves as a phonemic parameter in tone languages.

To test whether manipulated auditory pitch feedback would have the same effect on speakers of tone languages, Mandarin Chinese speakers were exposed to the same paradigm in a recent experiment [16]. Counter to expectation, the compensation and adaptation in the Mandarin speakers were similar to those observed for the English speakers. This is taken as evidence that postural and phonemic effects on \( F_0 \) control involve both internal representations and closed-loop auditory feedback. The speakers appear to have learned new internal models for \( F_0 \) control, irrespective of the particular linguistic function of \( F_0 \).

Thus, whereas the experimental evidence seems to support the distinction between phonemic and postural settings, it also

\(^1\) Notice, however, that pitch supports stress in English, as it does in many languages.
suggests that modifications of the auditory feedback signal have an effect on both types of parameters and that control of $F_0$ during speech production relies on both internal models and feedback-based adjustment.

3.2. Amplitude features
Auditory feedback is indispensable for proper control of overall speaking volume (SPL) in general, and vowel SPL in particular. In experiments in which a cochlear implant user’s speech processor was switched on or off, the speaker’s vowel SPL (and duration) changed in the first utterance after the switch occurred; the same effect was observed with normal-hearing subjects when auditory feedback via headphones was masked by noise [6]. One surprising result of these experiments was that segmental phonemic (vowel) contrasts may be affected as rapidly as vowel SPL.

Loud speech has been characterized as speech under the influence of natural perturbation [17]. Experiments with artificially perturbed speech, for instance bite block experiments, show that phonetic target regions may be reached by compensatory articulation strategies. In the loud speech condition, jaw movement and higher $F_0$ contribute to the perception of increased loudness. Since the perception of vowel height relies on the difference between $F_1$ and $F_0$, higher $F_0$ must be compensated for by higher $F_1$ to preserve vowel quality. Compensation strategies in loud speech thus involve interdependent articulatory and acoustic parameters [18].

3.3. Temporal features
Prosodic breaks subserve the structuring of utterances. The placement of breaks and pauses is affected by factors such as speaking style and speaking rate. Results of a study on the number and distribution of breaks as a function of different speech tempi suggest that different speakers may vary with respect to the implementation of rate effects [19].

In a follow-up study [20] it was found that a gradual parameter, which applies constraints on the length of intonational phrases, can account for speaking rate effects on the number and distribution of phrase breaks. However, if the speaker is introduced as an independent variable, the parameter range converges to one prototypical value for each speaker and speaking rate (here: normal, fast, slow). Furthermore, the study also showed that speaking rate in read speech is in turn partially a function of text genre.

It thus appears that speech rhythm and its variation is largely controlled by factors that affect postural settings (but see section 4).

3.4. Independent factors
The discussion of phonemic and postural effects on prosodic features suggests that these effects are triggered by a number of external factors. More concretely, we identify as independent factors: (a) speaking style as a communicative and situative factor, and (b) the acoustic conditions as another situative factor. As a first approximation we assume that speaking style tends to exert both phonemic and postural effects on prosodic parameters, whereas changes in the acoustic conditions mainly call for a continuous adjustment of postural settings.

In a review of work on prosodic cues that differentiate speaking style, speaking rate is listed among the most salient cues (others being the distribution of boundary tones and the rate of disfluencies) [21]. In particular, speaking rate is a good differentiator of read vs. spontaneous speech, being significantly faster in read speech.

Speaking rate in turn affects segmental and prosodic properties of speech. In the segmental domain, changes in speaking rate are known to have differential effects on the production of vowels as opposed to consonants, indicating different control strategies for the two types of speech sounds [22]. In the prosodic domain, speaking rate has been shown to influence the number and distribution of phrase breaks [19, 20]. The surface realization of accents and tones is also affected: the pertinent $F_0$ contours may be compressed or truncated in fast speech [23, 24].

4. Multilayered Models?
In this section we want to introduce two hypotheses: first, that the relative magnitude of phonemic and postural effects, respectively, on a given prosodic feature may be flexible, and second, that there may be more than one type of internal model that the speaker may rely on.

The acoustic correlates of prosodic features, viz. $F_0$, amplitude, and those pertaining to speech timing, can be regarded as variables which depend on a number of factors (see section 3). The factors can be characterized as communicative and situative settings, comprising the acoustic conditions in which the utterance is produced as well as different types of speaking style, including socially driven styles, situation specific styles, reading styles, and emotional styles [25].

During speech production, the speaker must implement the confounding effects of these factors on each acoustic variable. Moreover, the experiments with manipulated auditory pitch feedback [16] indicate that both internal representations and closed-loop auditory feedback are consulted for proper control of $F_0$.

Generalizing such observations, we suggest that the relative importance of acquired internal models of phonemic targets, on the one hand, and of immediate adjustments of postural settings, on the other hand, is flexible and depends on the actual communicative and situative conditions.

We further hypothesize the existence of more than one level of learned internal representations. It might be conceivable that the speaker may have acquired several models, each of which represents the most appropriate balance of phonemic and postural settings for a prototypical communicative and situative context. For instance, if the context calls for the production of loud speech, the speaker may access a prefabricated model that implements the appropriate compensation strategies in the articulatory and acoustic domains. When required to produce fast speech, the speaker may apply a different model, for instance one that reduces the number of phrase breaks, changes the surface realization of accents from complex to simple tones and truncates $F_0$ contours in certain syllabic and segmental structures.

5. Conclusions
A complete speech production model must incorporate segmental and prosodic properties of speech. 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) [26, 27]. There is experi-
mental evidence that speech production is planned with reference to prosodic structure (cf. the prosodic planning hypothesis [28]).

When compared to segmental characteristics of speech, which are best preserved by strong and stable internal representations [3], prosodic properties may rely more strongly on a balanced mixture of continuous, auditory feedback-based updates and learned internal models [16]. Based on evidence reported in the literature and on theoretical considerations we have presented two hypotheses: first, that the relative importance of acquired internal models of phonemic targets, on the one hand, and of immediate adjustments of postural settings, on the other hand, is flexible and depends on the actual communicative and situative conditions; and second, that the speaker may have access to several internal models, each representing the most appropriate balance of phonemic and postural settings for a prototypical communicative and situative context.

6. References


