

Modelling stylistic variation of speech

Basic research and speech technology application

Klaus J. Kohler

Institute of Phonetics and Digital Speech Processing (IPDS)
Christian-Albrechts-University, Kiel, Germany

ABSTRACT

Among the sources of pronunciation variability, phrase structure, prosody, communicative situation and speaking style are of prime importance for the modelling of connected-speech production. There is a need for complementing word-level phonetics and phonology by a phrase-level component through the analysis of large labelled connected-speech databases. This leads to the modelling of selected areas of phrase-level processes in basic research and in speech technology applications, especially in TTS synthesis.

1 Introduction

Pronunciation variability has 7 main sources:

- (a) regional and social variation between dialects and accents of the same standard language
- (b) phrase structure and prosody - e.g., conditioned by word class, position of words in sentence structures and utterances, stress, rhythm, intonation
- (c) communicative situation and speaking style
- (d) emotional expression
- (e) speaker state - e.g. fatigue, stress
- (f) gender and speaker type
- (g) the individual speaker.

This classification excludes temporary and permanent pathologies, e.g. drunken or Parkinson speech, as well as interactions between the variation classes. In all languages, phonetic research has primarily focussed on the pronunciation of words (in citation forms) under aspect (a), resulting, e.g., in pronunciation dictionaries and dialect atlases. Contextual variation under aspect (b) has been given less coverage, and connected speech variation under aspect (c) has only latterly received more detailed analysis. Areas (d) - (g) have been on the periphery of phonetic research, interest coming mainly from other disciplines, such as psychology and forensic studies. The main concern in the study of pronunciation variability has been with variation of phonetic segments or even simply of phonemes;

prosodic variability in its own right and as a conditioning factor for segmental variability has received far less attention until quite recently. And word-level phonetics and phonology has been over-emphasised at the cost of phrase-level phonetics, i.e. segmental variability in connected speech, especially in unscripted interaction.

In the study of areas (b) and (c), the phonological word-level approach, which establishes phonemic systems for word differentiation and allophonic variation controlled by coarticulation, is not sufficient. The phrase-level phonetic perspective, including prosody, is mandatory to observe, describe and explain, e.g., the processes of segmental assimilation, reduction and reinforcement in word concatenation. To investigate the statistical variation of these segmental processes large segmentally and prosodically labelled databases of connected speech are required, representing various speaking styles in a variety of scenarios.

Phrase-level phonetics in areas (b) and (c) has been a research focus at IPDS Kiel for many years [2,3]. It is now gaining momentum in a large number of speech centres for many different languages, as shown by the Sankelmark Conference *Patterns of Speech Sounds in Unscripted Communication* [5]. Its immediate goal is to derive sound patterns of connected speech for individual languages descriptively on the basis of statistical distributions of types of articulatory modification in large databases. These corpus analyses prove synchronic stylistic variation of the pronunciation of words to be regular and determined by a network of phonetic, prosodic, syntactic, semantic and pragmatic conditions. So the ultimate aim is to relate the statistical patternings found in individual languages to strategies in speech production, and to model stylistic variation at the phonetic phrase-level with reference to articulatory units and processes that enter into connected speech, and linguistic and communicative conditions that control their operation.

It is premature at this stage of our knowledge to attempt a comprehensive model for connected speech in any language, even the phonetically well described ones, such as English or German. But it is possible to offer general considerations for such modelling and to model specific areas of reduction phenomena language-

internally as well cross-linguistically. This paper aims to do such limited modelling for selected fields of stylistic variation in German corpora of spontaneous and read speech. Since areas (b) and (c) of pronunciation variability are also of great importance for the efficiency of ASR and automatic speech synthesis, TTS in particular, results of this basic research in connected speech modelling will be of special benefit for speech technology application. So this paper will finally point to some areas of TTS that may profit from an incorporation of connected speech model components, derived from German database analysis.

2 Methodology

The *Kiel Corpus of Read/Spontaneous Speech* [1] is a database for German that meets the conditions for the study of pronunciation variability sketched briefly in the Introduction. As a useful heuristic device for data retrieval and grouping, it relates the phonetic manifestations of words found in the corpus to canonical phonological forms through a structured sound alphabet and systematic labelling conventions. The structural basis of this alphabet is mainly segmental and quasi-phonemic, and the systematic labelling conventions include the marking of sound deletions, changes and insertions in relation to canonical symbols [4]. The transcription is thus of a broad contrastive type, omitting phonetic detail from preprocessing of the database at the symbolization level and leaving it to later database analysis. The broad transcription guarantees greater consistency in labelling by different labellers, on the one hand, and is still an effective means of accessing classes of phonetic forms for each word and classifying phrase-level processes symbolically for subsequent acoustic analysis. In addition, prosodic categories of word stress, sentence accent, intonation and prosodic phrasing are labelled with the labelling system PROLAB, based on the *Kiel Intonation Model* (KIM).

Spontaneous speech data show that sound segments may get deleted as far as segmentable units are concerned but still leave contrastive residues of phonetic properties in the environment, such as vowel nasalization linked to nasal consonant elision, or glottalization instead of plosives in the surrounding sonorant context. To cope with these non-segmental articulatory components the transcription alphabet of the *Kiel Corpus* also includes markers that are not given durations but are only allocated to points in time in the vicinity of which their phonetic exponents are to be found. These are effective symbolization devices for analysing phonetic processes such as nasalization, glottalization, palatalization linked to segmental reduction, rather than simple contextual coarticulation.

It is, however, crucial for the statistical corpus studies of transcriptions to transcend the phonetic symbol

strings in order to get to the articulatory processes and their linguistic and communicative embedding, which transcriptions only map in an abstraction from time and space. Symbols should not just be counted in various classes but need to be interpreted in relation to the speech producing mechanisms in linguistic structures and communicative settings, leading to a model of connected-speech production.

3 Modelling connected-speech production

3.1 General aspects: conditions, units, processes

Phrase-level phonetics is shaped by relations between

- the speaker's tendency to reduce effort
- the listener's demands for distinctivity
- the distinctivity requirements set by the communicative situation
- paradigmatic and syntagmatic linguistic patterns
- social conventions.

Paradigmatic differentiation of linguistic units influences the degree of reduction according to word class, for instance function vs. content word, and morphological category within function words (e.g. German *ihr* [ʔi:r] 'you(r)', when unstressed, can be reduced, at a low stylistic level, to [ɐ] as a personal, but not as a possessive pronoun). Syntagmatic structure in linguistic messages operates at a hierarchy of levels from syllables to words to morphological and syntactic constructions to semantic organization, and to prosodic grouping by accent, intonation and phrase boundaries. The prosodic features may support or cut across, any of the former syntagmatic elements. These groupings are marked by internal cohesion and junctural separation at the boundaries. Internal cohesion raises the probability of phonetic fusion inside the various syntagmas, whereas their boundaries have a high probability of being signalled by phonetic separators (e.g. the reduction of the German personal pronoun *ihr* to [ɐ] is only possible verb enclitically, not proclitically). The cohesive units are flexible: sequences of words may fuse to new lexical items (e.g. German *zum* [tsum] < *zu dem* [tsu dem]) 'to the', with separate semantics). Thus, in special cases, lexicalization has to be considered besides statistical variation of phonetic forms.

The effects of the 5 conditioning factors on phrase-level phonetic output cannot be deterministic, but lead to statistically variable reduction patterns. Different degrees of reduction can be arranged along an articulatory scale from canonical phonological form to extreme simplification, as in the German PREPOSITION + DEFINITE ARTICLE *mit dem* 'with the':

[mɪt^h d̥e:m] [mɪt d̥em] [mɪd̥em] [mɪd̥əm] [mɪbm] [mm].

The linear segmental representation of these forms hides the temporal patterns of coordinated articulatory parameters and creates the false picture of discrete segmental changes. The segmental phonetic transcriptions need to be translated into the timing of *supra-segmental* articulatory patterns, and it is these superordinated speech gestures that are changed in synchronization and composition of their articulatory parameters. These more global units are defined by the *opening-closing gestures* of the vocal tract: articulated speech is composed of sequences of these. It is thus not necessary to specify all the possible types of assimilation and reduction for individual segmental sequences, but quite general instructions suffice to trigger the application of a reduction program to various types of opening-closing gestures.

It may be assumed that speakers start from canonical forms, stored in a mental lexicon as sequences of opening-closing gestures with temporal information. In applying a reduction program to these forms a speaker has to set a value that tells the generation system what degree of reduction it is to achieve. This value is determined by the 5 conditions. So the system requires a *reduction coefficient*, operating at a high processing level before actual articulatory execution.

3.2 Example areas from German

3.2.1 Plosive + schwa + nasal syllables:

The realization without [ə] in the canonical poststress sequence PLOSIVE + [ə] + APICAL NASAL occurs with great regularity after fortis and lenis plosives in both read and spontaneous speech, but slightly more frequently in the latter. Left-to-right place assimilation of the apical nasal after labial/dorsal is equally frequent in both speaking styles, again a little less in read speech. These data suggest that schwa-less and at the same time place-assimilated forms have become the canonical lexical entries for the speaker group as a whole; the presence of schwa is a reinforcement, typical in the more formal reading style.

Through historical sound change, the opening-closing movement was optimally simplified, from a gestural point of view, by the elimination of the opening and by restricting the gesture to one oral articulator. This reorganization affected the opening-closing gesture as a whole by the application of a high reduction coefficient. But the reduction coefficient has remained 0 when the schwa syllable is not word final after stress. On the other hand, the reduction coefficient may be raised in the case of lenis plosive syllables to equalize velic lowering across the PLOSIVE + NASAL sequence, resulting in complete gestural integration. This is far less common than levelling of OPENING + PLACE and quite rare in read speech. It is most frequent with labial (due to the very frequent function word *haben* ‘have’), and least likely with apical outside function words (*werden* ‘will’ etc.). It occurs in the fortis context only in weak

prosodic environments, e.g. in unstressed *-zehnten* ‘-teenth’ of compound ordinal numerals or in unaccented *guten* ‘good’ of stereotype greetings.

3.2.2 Heavy syllables: the case of *-igsten*:

In all the *-igsten* words in the spontaneous speech corpus (inflected ordinal numerals and superlatives), the complex unstressed gesture opens from an apical stricture (mostly fricative) into a close palatal vowel. The canonical form then requires a closing into a palatal fricative stricture, followed by, and partially intertwined with, a closing into an apical fricative stricture. In more than 50% of the tokens the dorsal raising to fricative stricture is not carried out, thus making the global articulator movement between apical onset and offset more homogeneous. The simplification of the complex opening-closing gesture may be carried further by omitting the opening into a high vowel position, only leaving [s:t, st, s] in a small number of cases.

In the subsequent gesture [tən], the complete plosive occlusion is not established in ca. 10% of tokens. An opening into the vowel is not performed in 99% of cases (see 3.2.1), irrespective of the presence of a stop. All CONS + *igsten* occurrences can be linked to a sequence of two articulatory opening-closing gestures [n/t/sɪçst] and [tən]. The large spectrum of phonetic variants is then derived on the basis of the definitions of the gesture characteristics and assumptions about gesture reorganization. As the gestures are in unstressed syllables and are therefore executed with reduced articulatory energy they tend to have their complexities removed by successively raised reduction coefficients decreasing the extension of the movement as well as the number of participating articulators. The most extreme gesture reduction possible under these constellations is found in *vierundzwanzigsten nehmen* [fɪrən,svans ,ne:m] ‘take the 24th’ (g145a013).

3.2.3 Function words: Function words are by default unstressed, with higher reduction coefficients, than content words under identical conditions, affecting all articulatory parameters. Final clusters of CONS + [t], or intial [ts] (*ist* ‘is’, *nicht* ‘not’, *und* ‘and’; *zu(m)*) lose their stops. Vowels are reduced in duration and in quality, going as far as [ə] in, e.g., the indefinite article *ein*), and [ə] may get elided altogether. These durational and qualitative reductions are very important for stress-timing rhythm.

4 TTS application

In speech technology applications, ASR is confronted with the complete array (a) - (g) of pronunciation variability, consequently the demands on a universally usable and efficiently operating system are extremely high. But just taking the entire spectrum of possible reduction in areas (b) and (c) into account places

enormous demands on the efficiency of such systems, which cannot be solved without building in situational, semantic and syntactic constraints. TTS is more fortunate because the areas (d) - (g) can be ignored or standardized. Standardization also applies to (a). For (b) and (c), however, a certain amount of variability must be implemented in accordance with prosodic, syntactic, semantic and situational conditions, if the result is to be intelligible and to sound natural.

4.1 Formant vs. concatenative synthesis

Automatic speech synthesis has been dominated by the segmental approach (but cf. *YorkTalk* [6]), based on phonemic abstraction and correction through coarticulation. Formant synthesis can go a long way in efficiently implementing the interactive timing of vocal tract parameters, which is rather crudely captured by the concept of segmental coarticulation. If it still does not sound natural it has largely to do with the wrong, non-human voice source. Concatenative synthesis was not prepared to wait for phoneticians to come up with solutions and took engineering shortcuts by trying to capture natural sound production as well as mutual influences of contiguous segments on each other, i.e. coarticulation in the narrowest sense, in diphone units. The strategy was naive as it assumed that coarticulation was now contained in the units to be concatenated and no longer an issue to worry about. But coarticulation can have much longer extension, and diphone units change in different prosodic patterns. This led to working with demi-syllables, flexible multiphone units and even whole lexical items (e.g. strongly reduced function words). The output sounds natural because the units come from a human speaker. But the timing and the variability in connected speech are often wrong, because they need modelling of the type described in Sect. 3, beyond the simple concept of coarticulation.

Comparing the Infovox Desktop (concatenative) and the Infovox 230 (formant) TTS synthesis systems, this becomes very clear. The former has a natural, the latter an unnatural voice. But the temporal structure and the rhythm are better in the latter. Its rules, which are based on my work of about 10 years ago, reduce [əɪ] syllables as well as function words and modify abutting clusters at word boundaries to an extent that creates the correct timing and avoids accidental prominence on words that must be unaccented. The necessary temporal and spectral adjustment of concatenative units under varying prosodic conditions is not always handled adequately in these cases of obligatory reduction: all [əɪ] syllables have rather long vowels, and aspiration after preceding fortis plosives, and function words are often realised as phonetically too "strong". The data and modelling presented in this paper should help to develop new strategies for more efficient and natural sounding automatic (especially TTS) synthesis.

I have also continued working privately on formant syn-

thesis within the Rulsys/Infovox framework over the years and produced a version that goes well beyond the 230 system, in that it incorporates various speed and concomitant reduction controls, which activate whole packages of prosodic and segmental adjustments for a particular setting. This can be elaborated further, also to quantify the concept of a reduction coefficient.

4.2 Choice of speech level and priorities for implementing reductions

In automatic speech synthesis the stylistic level can be uniformly set (within an over-all speed and linguistic and prosodic constraints). This level will, for all usual applications, be a reading style, which means that extreme reductions, as found in a spontaneous speech corpus, are ruled out, but so are citation form pronunciations. What has to be given top priority in the implementation of German pronunciation variability in a TTS system are middle-range function word reductions (e.g. [mɪdəm]) as well as schwa syllable and consonant cluster (including geminate) simplifications. Further refinements can be introduced conforming to the data structures in the read corpus.

5 Outlook

The framework presented for the analysis and modelling of German phrase-level phonetics needs to be applied systematically to many languages. This research will open up interlanguage comparisons of connected speech processes, unravel common and divergent patterns, contribute to the study of typologies and universals above the level of phonemic systems, and assist multilingual speech technology applications.

REFERENCES

- [1] IPDS. *The Kiel Corpus of Read/Spontaneous Speech*. CDROMs #1-4, Kiel: IPDS, 1994-1997.
- [2] K. J. Kohler. "Segmental reduction in connected speech in German", in *Speech Production and Speech Modelling*, W. J. Hardcastle, A. Marchal (Eds.), pp. 69-92. Dordrecht: Kluwer, 1990.
- [3] K. J. Kohler (Ed.). *Sound Patterns in German Read and Spontaneous Speech*. AIPUK vol. 35, Kiel: IPDS, 2001.
- [4] K. J. Kohler, M. Pätzold, A. P. Simpson. *From Scenario to Segment*. AIPUK vol. 29, Kiel: IPDS, 1995.
- [5] K. J. Kohler, A. P. Simpson (Eds.). *Patterns of Speech Sounds in Unscripted Communication*. J of Inter Phon Ass 31(1), Cambridge: CUP, 2001.
- [6] J. Local, R. Ogden. "A model of timing for non-segmental phonological structure", in *Progress in Speech Synthesis*, J. P. H. van Santen, R. W. Sproat, J. P. Olive, J. Hirschberg (Eds.), pp. 109-121. New York: Springer, 1997.