

Investigating unscripted speech: Implications for phonetics and phonology

K. J. Kohler

Abstract

This paper looks at patterns of reduction and elaboration in speech production, taking the phenomenon of plosive-related glottalization in German spontaneous speech, on the basis of the 'Kiel Corpus', as its point of departure, and proposes general principles of human speech to explain them. This is followed by an enquiry into the nature of a production-perception link, based on complementary data from perceptual experiments. A hypothesis is put forward as to how listeners cope with the enormous phonetic variability of spoken language and how this ability may be acquired. Finally, the need for a new paradigm of phonetic analysis and phonological systematization is stressed, as a prerequisite to dealing adequately and in an insightful way with the production and perception of spontaneous speech.

1. INTRODUCTION

The phonetic question to be discussed in this paper may be highlighted with a joke by the British comedians Ronnie Barker and Ronnie Corbett from one of their television shows [Davidson and Vincent, 1978, p. 142]: "Four girls were disqualified for cheating in the Miss Greater Manchester competition last night. They were Miss Altrincham, Miss Paddingham, Miss Pumpingham and Miss Stuffingham."

For those readers who are not too familiar with English geography and with the intricacies of the spelling-sound relationship of English place names, "Altrincham" is a town south of Manchester, pronounced [ˈɒltrɪŋəm], just like "Birmingham" or "Nottingham". The places the other three beauty competitors are suggested to come from do not exist but are given names that follow the same pattern of name formation in "-ham" [əm]. Over and above that, the real and the pseudo place names are also interpretable as phrasal constructions of verb + "them" with articulatory reduction at the sentence and utterance levels. The two Ronnies play on the linguistic indeterminacy of speech signals, which results from the enormous flexibility of word production in utterance context, irrespective of regional and social variation of a language. Their word play builds and depends on the listener's intuitive knowledge of a many-to-one relationship between word sequences and their acoustic output in speech production, on the one hand, and of a one-to-many relationship between acoustic speech signals and units of meaning in speech perception and comprehension, on the other. Without the phonetic variability and semantic indeterminacy as essential constituents of speech communication, jokes of this kind, examples of which may be multiplied across languages, would fall flat.

The question then arises as to how phoneticians and phonologists can cope with these facts in their modelling of speech communication. To give an adequate answer they should realise that humans communicate, first and foremost, in unscripted interaction and that, to gain insight into speech production and perception processes as well as into speech development, the variability of unscripted speech should therefore be a focus of attention. They must also be aware that this variability goes far beyond allophonic variation of segmental-type phonemes in word pronunciations; it includes, amongst others, the phonic flexibility of speaking styles and adjustments to the communicative situation. However, phonetic analysis and phonological systematization have not paid due tribute to this prime importance of unscripted speech communication. They have instead concentrated on lab, or at best textual, speech frames for the production and perception of sounds in words, rather than of words in utterances, and they have done this with a view to setting up **patterns of rigid invariance in word phonology** instead of providing **rules of structured variability in utterance phonology**.

The latter approach is needed for an explication of everyday speech communication. Utterances set frames for phonetic flexibility of words in speech production, and words require utterance embedding to be perceived and understood appropriately. On the one hand, this utterance dependence controls the phonetic coalescence and semantic ambiguity in context, as is exemplified in the joke, and on the other hand, it is responsible for extreme articulatory reduction but nevertheless correct decoding in context, as is shown by countless instances of spontaneous speech. Remove the utterance contexts, and the remaining word sequences become unintelligible. This may be illustrated by the following example from the 'Kiel Corpus of Spontaneous Speech' [IPDS, 1995, 1996, 1997]: "nun wollen wir mal kucken" ("*now let's see*", g122a009) in the phonetic form [nũ: ð̃ᵐ ẽ̃ŋa 'kʰʊkŋ] for unreduced [nu:n vɔləŋ vi:rə mal 'kʰʊkŋ]. It has strong

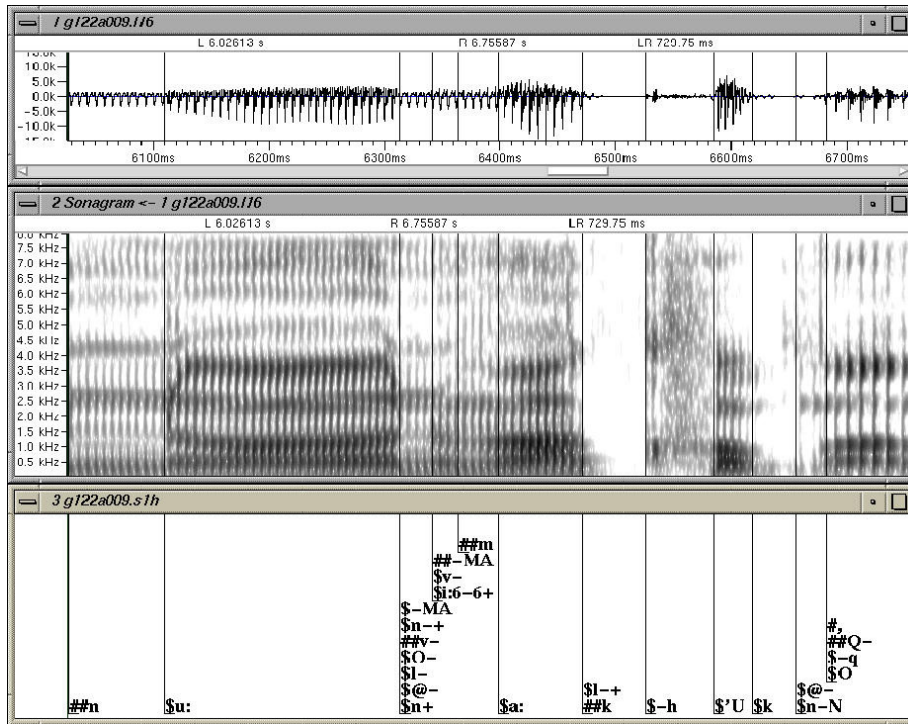


Figure 1
Speech wave, spectrogram and SAMPA label window for the 'Kiel Corpus of Spontaneous Speech' utterance "nun wollen wir mal kucken" ("now let's see") in dialogue turn g122a009: reduced speech.

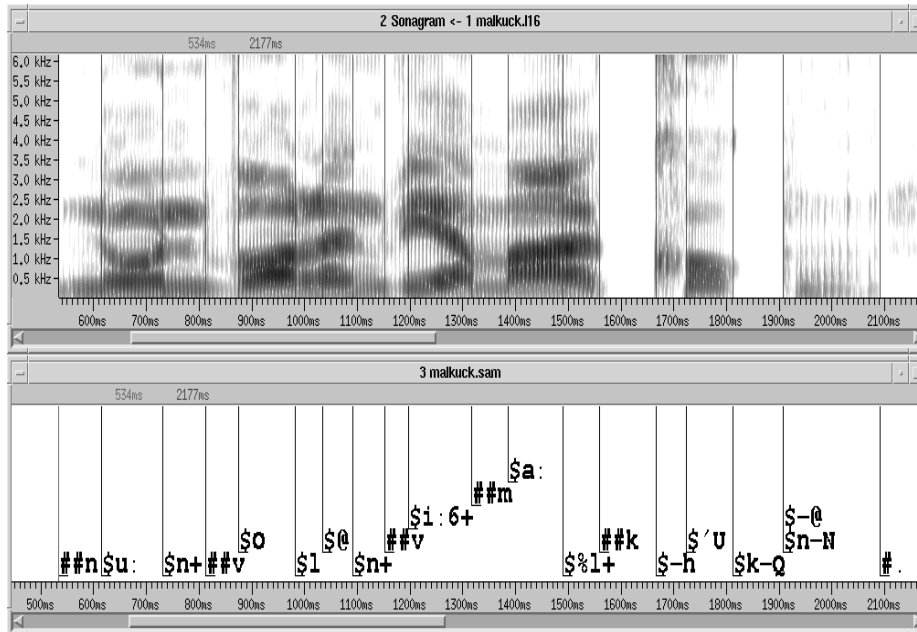


Figure 2
Speech wave, spectrogram and SAMPA label window for a reading pronunciation of "nun wollen wir ma kucken" by speaker KJK: elaborated speech.

nasalization across its first three syllables relating to syllable-final nasal consonants, which are reduced (deleted or shortened) in this hypo as against the hyper pronunciation. There is additional labiodentalization around the third syllable representing canonical [v] of "wir". Other possible realizations are [nū: ɔ̃ŋ/mě ma 'kʰʊkŋ]/[nū: ɔ̃ě ma 'kʰʊkŋ], where the apical gesture of the

medial nasal is also eliminated or the consonant deleted altogether. A native speaker of German has no problem in understanding the whole utterance, but when "kucken" is removed, decoding the larger section that remains becomes impossible. The spectrograms of figures 1 and 2 compare the spontaneous speech manifestation of this sentence with a careful reading pronunciation, contrasting the acoustic consequences of levelled versus extensive articulatory movements.

This paper aims to correct the traditional paradigm of phonetic/phonological sound segment and word orientation, taking a large corpus of spontaneous German dialogue data ('The Kiel Corpus of Spontaneous Speech') as its point of departure. Four questions will be addressed:

- The first deals with data and theory of **speech production**:
What are the patterns of reduction and elaboration in speech production and what general principles, governing human speech, can be adduced to explain them?
- The next two deal with data and theory of **speech perception** as well as with **speech development**:
(a) What role do the production phenomena play in perception and what is the production-perception link?
(b) How do communicators succeed in relating a large array of phonetic forms to the "same" item – a word or an utterance, and how may this ability be acquired?
- The fourth question deals with a **new paradigm for phonetics and phonology**:
What categories of description are necessary to systematize the variability of speech for an adequate and insightful account of its production and perception?

The discussion will pick out a frequently occurring phenomenon of unscripted German speech – glottalization in connection with plosive articulation – and interpret findings from corpus data on speech production and from experimental data on speech perception.

2. GLOTTALIZATION IN THE PRODUCTION OF PLOSIVES

2.1 Survey of data on stop production in German

In German connected speech – text reading and especially spontaneous dialogues – glottalization, in alternation with, or in addition to, more forceful glottal stops is a very common phenomenon. It not only applies to the context of word-initial vowels, which has been acknowledged in textbooks for a long time, but also to two further contexts:

- (1) 'sonorant - plosive - sonorant' (especially nasal) for fortis as well as lenis stops at all places of articulation, e.g. "könnten" [kœnn̥n] or "Stunden" [ʃtunn̥n] or "sind noch" [zɪn̥ nɔx], instead of the more elaborate canonical pronunciations [kœnt^hən], [ʃtundən] (or [kœntn], [ʃtundn] with nasal plosion), [zɪnt nɔx]; Figure 3 contrasts the reduced forms of "können" [kœnn], with a modal-voice nasal, and "könnten" [kœnn̥n], with a glottalized nasal;
- (2) 'vowel - fortis plosive - consonant (especially nasal), e.g. "zweiten" [tsvai̯(?)n], "Leipzig" [laɪ̯ʔptsɪç], "hat nicht" [hɑ̯ʔ nɪç].

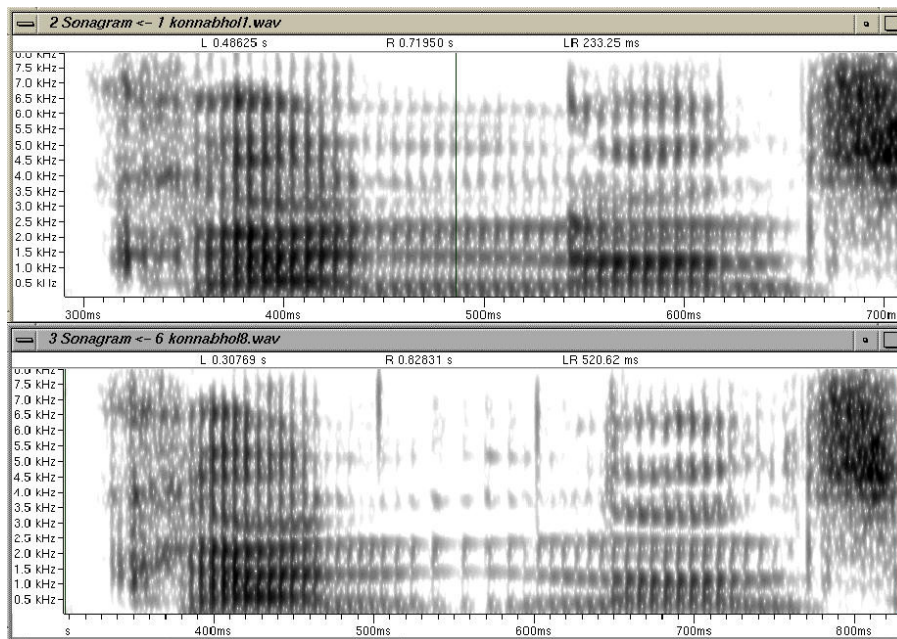


Figure 3
Spectrogram of "(Die) können uns (abholen)" ("*They can collect us*") (top) and of "(Die) könnten uns (abholen)" ("*They could collect us*") (bottom); read speech, speaker KJK.

Both these types of glottalization are related to plosives in more elaborate, especially citation form renderings of the relevant words, without or with intervening morpheme or word boundaries after the plosive. In these cases a simple glottal valve action is used to cut off the air stream for stop articulation. This may happen in addition to or instead of a more complex combination of supraglottal oral and velic closures. A single long hold for a glottal stop may also be relaxed into irregular pulsing of rather long periods, compared to the environment. Moreover, the stop is not released into a vowel but in most cases is followed by another complete or partial oral occlusion – nasal, plosive or lateral. This articulatory sequencing is usually the result of [ə]-elision before nasals or laterals of canonical forms. In the complete nasal environment of case (1), the oral closure may be at the same place of articulation, and be accompanied by velic opening, throughout the sequence, as the interruption of the air stream is transferred to the glottal valve.

Under condition (1) there are four possibilities of temporal alignment of glottalization with the sonorant: medial, final, initial or complete irregular voicing, i.e. (a) [n̥n̥], (b) [n̥n̥], (c) [n̥n̥], (d) [n̥n̥] for apical nasal articulation in, e.g., "können". The following distributions have been found for the four categories:

- (a) is by far the most common in all contexts,
- for fortis stops (c) is the next frequent,
- for lenis stops it is (b).

In the context of condition (1) there are also occurrences of voiceless nasals instead of fortis, and of breathy-voiced or voiceless nasals instead of lenis stops, due to glottal (interarytenoidal) opening. In all these cases the modal-voice nasal context is interrupted by a different type of phonation as a residue of more complex plosive articulations. But the lenis context also allows a further progression to modal voice, i.e. reduction to [nn]. In the fortis context this is only possible in unstressed function words and elements of compounds, e.g. "-zehnten" in numerals. This

change may be complete, or there may be a very weak trace of the plosive in the form of a medial amplitude and/or F0 dip in the nasal stretch. So the speaker still signals a break, albeit towards the low effort end of a reduction scale ranging from plosive to complete nasalization (for further details see Kohler, 1996a, 1998).

2.2 Explaining the data with reference to general principles of speech production

All the phonetic realizations in nasal contexts of (1) eliminate the need for a synchronization of velic control and increase coarticulatory ease through a transfer of the valve action from the velum to the glottis. The velum can thus remain lowered in the entire sequence, but for canonical fortis plosives and the majority of lenis ones the listener is still guaranteed a signal break through a glottal stop, glottalization or some other change in phonation. These articulatory reductions at the utterance level may be regarded as instances of a more economical articulatory reorganization under the general principle of economy of effort, even if it is at present not quantifiable. In the typical nasal plosion context of condition (2), glottal compression is also possible although less frequent than in (1). Since the synchronization of velic control is then again no longer crucial this glottal adjustment also constitutes an increase of coarticulatory ease through reorganization.

Glottal stop and more relaxed glottalization can thus accompany plosive articulations in non-aspirated environments, which require adducted, rather than abducted vocal folds. In nasal contexts a scale of glottal phenomena can take over the stop function altogether and thus allow velic action and velic synchronization, i.e. movements of a sluggish articulator, to be eliminated. Moreover, the timing of these phonation changes within the nasal stretch can be left indeterminate, as long as they occur. That reduces the demands on articulator coordination and helps to ease production whenever called for by the context of situation.

3. PLOSIVE-RELATED GLOTTALIZATION IN PERCEPTION

3.1 A hypothesis

Since the production patterns of plosive-related glottalization are wide-spread, rule-governed and referable to general principles of speech production it must be assumed that their acoustic manifestations play a fundamental role in speech perception. The question thus is in what way listeners make use of, e.g., the presence/absence of glottalization and of its temporal indeterminacy to restore the intended utterances containing words with or without stops.

3.2 A perception experiment

For this purpose the utterance "die können uns abholen" ("*they can collect us*") of Figure 3a was used as the base for stimulus generation for a perceptual experiment. About 65ms of glottalization from another utterance "die könnten uns abholen" ("*they could collect us*") of Figure 3b were spliced into the long nasal of "können", replacing its initial, its central or its final section or (by doubling) the entire length. Figure 4 shows the speech waves of original "können" and of the 4 glottal splicings. Furthermore, the modal nasal of "können" was lengthened and shortened in another two stimuli. Finally both original stimuli together with the 6 manipulated ones were duplicated 10 times, randomized and presented to 23 subjects in a formal listening test

for identification of the test utterance as containing either the word "können" or the word "könnten".

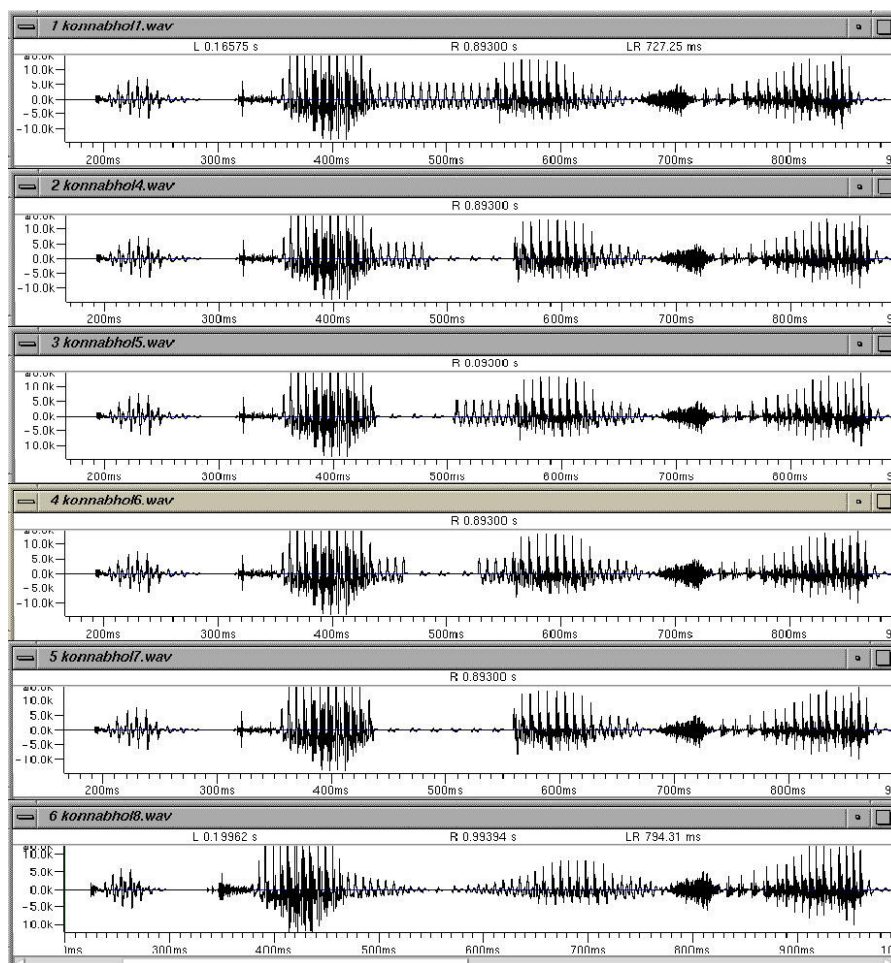


Figure 4
Speech waves of original "Die können uns ab(holen)"(top), "Die könnten uns ab(holen)"(bottom), and 4 stimuli derived from "Die können uns ab(holen)" by splicing, with glottalization replacing the second half, the first half, the centre, or the total length of the nasal, respectively. Original stimuli as in Figure 3. Stimuli generated for perception experiment.

The results are very clear: the presence of glottalization produces practically 100% "könnten", its absence 100% "können" judgements. So the presence as well as the temporal indeterminacy of a glottalized section in production are mapped onto perception. The listener decodes the break in modal voicing of the nasal – at least as long as it is of a duration typically found in production – as a cue to a stop, and ignores the precise synchronization with the nasal, in the same way as the speaker differentiates between presence or absence of a stop (for further details see Kohler, 1999).

3.3 Prolegomena to a perception theory for speech communication

What type of speech perception theory can account for this mapping of stop articulations? The Motor Theory or Direct Realism are most unlikely candidates because there is no invariant articulatory gesture that may be said to underlie the large variability in production for the listener to recover. What happens on the part of the speaker is a short-term interference with the air

stream, which in the extreme case is a complete stoppage of airflow out of mouth and nose. This essential aerodynamic goal may be achieved by widely differing articulatory gestures: a simple glottal or a more complex supraglottal occlusion or both. The glottal obstruction may be relaxed, resulting in glottalization and other phonation types that deviate from the surrounding modal voice, even if only by a decrease in amplitude and/or fundamental frequency. They thus still mark the intended goal of local air stream interference. What mechanism is actually used by the speaker depends on the amount of effort that is to be put into the production of speech as a function of utterance context and communicative situation. If the supraglottal occlusions are relaxed the closure periods go to zero, which means nasalization in a nasal context and lenition in an intervocalic context. Both are wide-spread in the languages of the world. In the case of nasalization the short-term interference with the air stream may be removed altogether; similarly, intervocalic lenition can result in complete vocalic integration.

As long as air stream interference for intended stop articulation is produced by the speaker, in whatever way, it is mapped onto the acoustic signal as a short-term change from modal voice. This, in turn, is what listeners can rely on as a common feature, if they have learnt to group the variability in their phonetic experience through exemplar-based learning, in a way Björn Lindblom is proposing in his paper [Lindblom, this volume]. So the perception theory that is to cope with reduction phenomena in spontaneous speech successfully will have to have an acoustic and auditory base. And it can no longer postulate phonetic invariance of any sort, articulatory or otherwise, because the speech material the listener receives for decoding from speakers in everyday interactions is simply not made that way, but listeners still cope extremely well with this structured variability.

Constraints from higher levels of speech processing will also have to be integrated into such a perception theory for spontaneous speech. If German listeners receive an acoustic signal that does not contain a break in nasal modal voicing they will interpret it as having no stop component at the signal level, e.g. being "können" rather than "könnten" or the cardinal number "dreizehn" rather than the ordinal "dreizehnten". But since in unstressed function words and numeral components nasalization may also affect fortis stops (the 'Kiel Corpus' provides several instances of [tse:n(n)] "-zehnten", e.g. g125a005, g256a001), a stopless signal may actually refer to an intended conjunctive or ordinal form containing a stop. So "-zehnten" in "am dreizehnten November" ("*on November 13th*") and "-zehn" in "an dreizehn Novembertagen" ("*on 13 days in November*") may coalesce phonetically, but they will, of course, still be decoded as the ordinal and the cardinal number, respectively, due to top-down interpretation.

3.4 Some thoughts on speech and language development

The next question concerns how the link between the production of structured variability in utterances and their correct decoding develops in speech and language acquisition. I find Björn Lindblom's reference to the roles of motor constraints and of perceptual experience very attractive [Lindblom, this volume]. Projected onto our data his "low-energy articulatory search" would enable the child spontaneously to discover simple glottal valve patterns by the side of other more complex stop mechanisms, both used by the ambient phonology. The low-cost motor patterns of adult reduced speech would thus also be arrived at by the child in a playful exploration of basic air stream control. The child would thus gain phonetic experience with its own speech actions and their relationships from the energy angle, and it would, in parallel, gain experience with the acoustic patterns produced by adults, which show a certain degree of

congruence to the child's own in respect of low-cost stop control. It could thus be able to form "perceptual categories as emergents of phonetic experience". Research into child language should pursue these questions by looking at the development of phonation types, also in connection with stop articulations.

4. A NEW PARADIGM FOR PHONETICS AND PHONOLOGY

The data presented in this paper cast new light on theories of speech production and perception that are required for a modelling of the speech communication process. We need a new paradigm for phonetic analysis and phonological systematization as a prerequisite for such theory development [Kohler, 1996b]. The following five points summarise the essentials for such a new paradigm.

(i) Connected and especially unscripted speech must be given a much more prominent role in speech research than has been customary. We need speech data of at least sentence size in natural and meaningful contexts because it is only there that reductions occur and can be tested perceptually.

(ii) The focus of attention in speech production and perception is to be shifted from word to utterance phonology, from phonemes in words to phonetic shapes of words in utterances.

(iii) The search for invariants at any level of production or perception is to be given up in favour of the grouping of variants according to general principles of motor constraints and according to perceptual experience and exemplar-based phonetic memory. This approach emphasises the importance of signal statistics in category formation, which traditional phonology, in contrast to automatic speech recognition, has denied.

(iv) The strictly linear segmental phonemic frame for phonological systematization has to be complemented with nonlinear componential features referring to any articulatory or phonatory aspect. This is mandatory in view of the temporal indeterminacy that was found in plosive-related glottalization in German production and perception data: if a phonetic feature is not used in a linear segmental way by speakers and listeners, it should not be treated as such in phonological analysis. It is characteristic of reduced speech, compared to elaborated canonical phonetic forms, that the linear segmentability, which is customarily applied to the latter, becomes fuzzy: features of glottalization, nasalization, labi(odent)alization, palatalization, velarization etc. become dissociated from specific linear segments and temporally indeterminate, they operate as long components or articulatory prosodies instead [Firth, 1948].

This may be illustrated by further examples from German that combine several of these prosodies. "soll sie" [zɔzi] ("is she to") vs. "sollen sie" [zõzi] ("are they to") vs. [zõzi] "sollten sie" ("should they") – for canonical [zɔl zi] vs. [zɔln zi] vs. [zɔltn zi] – are differentiated by the presence or absence of nasalization and/or glottalization in the vowel of the first syllable, and the linear segments /l, t, n/ of the canonical forms are no longer discernible. These articulatory prosodies are also relevant for the listener as was shown by a formal listening test with these items in contextual frames: [zõzi] was predominantly decoded as "sollten sie" (for further details see Kohler, 1999).

Research into speech and language development in children would also benefit immensely if the

established frame of phoneme acquisition were given up and the development of children's long-component production patterns as well as the development of their more global perception patterns were studied instead. This would also assist the theory building sketched above.

(v) Finally, the time-honoured division of the field of phonetic science into phonetics, which provides raw measurement data, and phonology, which cooks them and turns them into a good linguistic meal, has outlived itself, although it is still prevalent in the heads of many linguists. This division advocates the supremacy of phonological categories and their independence as well as preexistence vis-à-vis phonetic substance and has led phonetic research astray on many an occasion in the past, e.g. when psychologists took over the phoneme concept and developed perception or language acquisition models, or when linguists created the sub-discipline of lab phonology and the model of articulatory phonology. The division has now reached the stage where it becomes a hindrance to the advance in the theory and modelling of speech communication as it unfolds in real speakers and hearers in natural settings every day and everywhere.

REFERENCES¹

(AIPUK = Arbeitsberichte d. Inst. f. Phonetik u. digitale Sprachverarbeitung d. Univ. Kiel)

Davidson, I.; Vincent, P.: *The Bumper Book of the Two Ronnies. The very best of the news.* (W.H. Allen, London 1978).

Firth, J. R.: *Sounds and prosodies.* Transactions of the Philological Society, 127-152 (1948).

IPDS: *The Kiel Corpus of Spontaneous Speech. Vols. 1-3. CD-ROM#2-4.* Institut für Phonetik und digitale Sprachverarbeitung, Kiel 1995-1997)

Kohler, K. J.: *Phonetic realization of German /ə/-syllables.* AIPUK 30: 159-194 (1996a).

Kohler, K. J.: *Developing a research paradigm for sound patterns of connected speech in the languages of the world.* AIPUK 31: 227-233 (1996b).

Kohler, K. J.: *The phonetic manifestation of words in spontaneous speech.* In Duez, D. (ed.), *Proc. from the Esca Workshop on Sound Patterns of Spontaneous Speech, La Baume-les-Aix: 13-22* (1998).

Kohler, K. J.: *Articulatory prosodies in German reduced speech.* Proc. 14th Inter. Congr. Phonet. Sci., San Francisco: 89-92 (1999).

Lindblom, B.: *Developmental origins of adult phonology. The interplay between phonetic emergents and the evolutionary adaptation of sound patterns.* *Phonetica* 57: 297-314 (2000).

¹Graphic signal representations and speech output of utterances referred to in this paper can be found at the following URL: http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2000_1/kk_99b.html