THE DOOR FRAME STRUCTURE OF SPEECH SCIENCES

The famous German DUDEN dictionary (Drosdowski, 1996) defines *complexity* by "multi-facetedness" and "the interplay of a large number of features". A counterpart of the DUDEN for English, the Oxford English Dictionary (Winchester, 2003), characterizes *complexity* as something that is "intricate" and for this reason often also tricky and messy. In other words, *complexity* is the result of:

- numerous *categories*,
- plus numerous (mutual) *relationships/links between* these categories,
- and a lot of *variation within* these categories.

It leaps to the eye that these three key features of *complexity* likewise describe the quintessence of language and speech communication. When we speak to each other, we have to convert our communicative intentions (which are already categories by themselves) over time into en- and decodable categories that range from numerous prosodic units like phrases, words, and syllables to individual sounds and melodic configurations at sentence accents and phrase boundaries. These larger and smaller "morsels" relate to each other in many different ways. For example, they are embedded in a recursive syntactic structure, they can function as differently weighted redundant features of the same communicative intention, and they influence each other during the processes of speech production and transmission. These processes – together with the trading relations of redundant features and biologically, situationally, and socially determined between-speaker differences – are at the same time also the major sources of variation within the larger and smaller "morsels" that constitute the speech code.

That is to say: *Complexity is inherent in speech*; and neither can we, nor do we, as speech scientists, want to reduce this complexity. Rather, we have learned to look at this complexity from various perspectives. We observe that language users tend to complain that their languages get poorer, often with reference to the way it is used by the next generation of speakers; and maybe nowadays these complaints are becoming louder and more frequent, since new digital communication channels, as well as their high transmission speed and availability, undoubtedly change

the ways we speak to each other. Against this background, it is our task as speech scientists to remind those worried people that it is the entirety of communicative intentions that determines the complexity of speech, and that the total number of intentions will not decrease and thus always place the same high demands on the number of en- and decodable categories that linguistic systems must provide to speakers, no matter how future channels and ways of communication may look like. Pellegrino et al. (2009: 2) call this the "balance of complexity"; in the words of Jakobson (1973: 48): "Like any other social modelling system tending to maintain its dynamic equilibrium, language ostensibly displays its self-regulating and self-steering properties".

In fact, linguistic systems are a bit like street and transportation systems of large cities. Some refurbishment works – which are indeed more often annoying than stunning – are inevitable every now and then in order to meet the requirements of new constructions and the next generation of users. But, if we decide to conduct roadwork or close down a metro line (just as we can decide to stop using a certain sound, word, or grammatical concept), we have to compensate for this loss in another place of the system, as the total number of travellers and packages that pass through the system remains the same.

Correspondingly, when we speak of "tackling the complexity of speech", as in the title of this volume, what we mean is that our chief aim as speech scientists is to analyze and organize the data that we collect "on the street", and on this basis to explain, understand, and model the speech patterns that we find. After the term "complexity" received little attention in the past of speech sciences, our volume is now already the second in a short period of time that explicitly takes up this term. The previous volume of Pellegrino et al. (2009) sets out ways of developing new phonology frameworks and predictions about language evolution by adopting key concepts of the general science of complexity: *self-organization*, *emergence*, and *nonlinearity*. There is of course some overlap with the content of our volume. However, first and foremost, we complement the perspective of Pellegrino et al. That is, in our volume, the complexity of speech is not the subject of research, but a given fact; and we show – from a phonetic rather than phonological point of view – how this fact shapes

the questions, methods, and findings that are around it and makes them increasingly more interdisciplinary.

The three actions of:
– collecting and analyzing data,
– explaining data,
– and organizing and modelling data.

constitute our door frame into the phenomena of speech. Data collection and analysis on the one hand and data organization and modelling on the other are the two vertical pillars of the door frame. They represent the main part of the work and provide the hinges on which the door is supported and turns. Explaining the data corresponds to the door lintel. It links the two vertical pillars and stabilizes the entire structure. This door frame metaphor provides two important insights.

First, tackling the complexity of speech involves two diametrically opposed efforts. We have to go on discovering the complexity of speech by piling up speech material as well as measurements and judgments. These efforts constantly *increase* empirical complexity. In parallel to this, we have to *manage and reduce* this empirical complexity again by developing models, representations, annotation systems, and the like. So, tackling the complexity of speech does not merely mean to take the current state of the art and try to simplify it. Rather, we have to acknowledge that we first of all have to make things more complex, before we can make sense of and simplify them.

Second, the opposed efforts of increasing and reducing empirical complexity and hence the two vertical pillars of the door frame metaphor are reminiscent of the differentiation between "phonetics" and "phonology", although neither phonetics nor phonology exactly corresponds to one of the two door frame pillars. For one thing, phonetics is of course much more than the collection and analysis of speech data (cf. Kohler, 1995, 2000). For another thing, Ohala (2004) differentiates between scientific and taxonomic phonetics and states that "phonology unquestionably embraces taxonomic phonetics" (p. 134). Nevertheless, the door frame metaphor nicely illustrates in this context that phonetics and phonology are indeed as closely linked as two sides of the same coin. It is important for speech scientists to bear this fact in

mind, not least because it was not always as self-evident as it is today (Ohala, 2010).

It was about hundred years ago that research into speech communication phenomena started using instrumental and experimental tools and techniques; and for some primarily phonetic researchers of those days, parameters and measurements became a replacement rather than a complement of traditional phonetic methods, the trained ear in particular. Sweet (1911) forcefully insisted in this context that "the final arbiter in all phonetic questions is the trained ear of a practical phonetician: differences which cannot be perceived must – or at least may be – ignored; what contradicts the trained ear cannot be accepted". Above and beyond Sweet's plea, a trained ear is not only important for determining the relevance of measurements. Trained ears also establish the decisive connection between measurements on the one hand and their interpretation in terms of communicative meanings and functions on the other, particularly when these trained ears come from a combination of native *and* non-native speakers of the corresponding language. That phonetics focused maybe too strongly on the tempting magic of numbers is probably one reason why phonology developed to some degree into the opposite direction, starting with Structuralism (with structures and binary oppositions already being in some contrast to gradual phonetic parameters) and culminating in the assumptions that phonology can largely do without detailed phonetic input, and/or that phonetics is just to fill phonological frameworks with some (messy and variable) surface substance.

By splitting up in the sketched ways, both phonetics and phonology lost, to a certain degree, their touch with the crucial lintel of the door into speech communication; or, in other words, what fell by the wayside to some degree during this period of "estrangement of phonetics and phonology" (Ohala 2004: 136) in the 20th century were the efforts to really explain and understand the complex patterns of speech, and how they relate to communicative meanings and functions. Fortunately, phonetics and phonology are growing together again in the 21st century, probably encouraged by the new challenges and opportunities of modern speech research, which cannot be met by primarily descriptive phonetic

analyses or phonological concepts and models that are not deeply rooted in phonetic data:

First, we have made enormous progress in the last decades in discovering the actual complexity of speech, for example, by studying melodic patterns, prominences, as well as types and levels of phrasing and voice quality on an equal footing with sound segments. This equal footing includes developing prosodic phonologies and linking them with existing segmental phonologies. At the same time, we have successively shifted our research focus from isolated words and sentences through read texts to real spontaneous everyday dialogues. Moreover, we have started filling and bridging the empirical gaps between segments and prosodies as well as between prosodies and gestures, and we have developed an awareness of the differences between speakers and registers and the strengths and pitfalls of speech elicitation and analysis methods (Niebuhr & Michaud, 2015).

Second, the great new challenges and opportunities for the modelling of speech mainly result from all the robotic, medical, and forensic applications that sneak into our everyday life in continually growing numbers. This includes technically supported ways of communication, speech processing tools, speech-based evaluation of technical systems, and human–robot interaction. Stunning progress has been made in each of these fields. Who would have thought a couple of years ago that finding out how robots can best socialize with humans (for example, without sounding either too dominant or incompetent) could ever be on a phonetician's agenda? In addition, progress in speech processing tools allows us to (semi-)automatically annotate and analyze amounts of speech data that the previous generations of speech scientists would have considered completely unmanageable or enough for a whole life.

Third, particularly in connection with the modelling of speech, recent progress in the organization of speech data led to the development of new forms and categories of representation and annotation for both segments and prosodies; and, as the new "approaches to phonological com-

plexity" of Pellegrino et al. (2009) suggest, thus development is far from complete. For example, the papers of Wagner et al. and Kügler et al., presented at the International Congress of Phonetic Sciences 2015 in Glasgow, Scotland, suggest new categories and concepts for a more fine-grained and comprehensive prosodic annotation. Recent developments in the organization of speech data also made us better understand which parameters of the speech signal are most fruitful and most reliable for annotating and analyzing speech corpora with respect to certain communicative meanings and functions. Moreover, some progress has even been made in measuring complexity: "concrete measures of complexity have been proposed or at least considered for features, segments, and syllables" (Chitoran & Cohn, 2009:39).

The challenges and opportunities of modern speech research place high demands on speech scientists, particularly those with a linguistic basis. In addition to being experts in the traditional key concepts, methods, and theories of both phonetics and phonology, they also need detailed insights and skills in many aspects of speech technology – for example, in order to create, search, and analyze their speech corpora – or collaborate with engineers and medical or social scientists on various applications. Those collaborations also require having a knack for financial and business matters, not least because acquiring research funding is currently more important than ever. The increasing interdisciplinary of the speech sciences (cf. Laver, 2001 for arguments in favour of the plural "sciences"), in combination with the rapid progress and the growing number of researchers in each field, make it successively harder for speech scientists to maintain an overview of the entire research on speech communication.

For this reason, the present volume is meant to give the reader an impression of the range of questions and topics that are currently subject of international research in:

- the discovery of complexity,
- the organization of complexity,
- and the modelling of complexity.

These are the main sections of our volume. Each section includes four carefully selected chapters. They deal with facets of speech production, speech acoustics, and/or speech perception or recognition, place them in an integrated phonetic-phonological perspective, and relate them in more or less explicit ways to aspects of speech technology. Therefore, we hope that this volume can help speech scientists with traditional training in phonetics and phonology to keep up with the latest developments in speech technology. In the opposite direction, speech researchers starting from a technological perspective will hopefully get inspired by reading about the questions, phenomena, and communicative functions that are currently addressed in phonetics and phonology. Either way, the future of speech research lies in international, interdisciplinary collaborations, and our volume is meant to reflect and facilitate such collaborations.

The volume section on the discovery of complexity begins with two chapters on speech prosody, which illustrates the great importance of this subject in current research. Churaňová addresses the understudied rhythm of Czech on the basis of speech-metronome synchronizations. She investigates experimentally how speech is produced – more specifically, how disyllabic words with different syllable structures are timed – relative to a constant metronome beat, and what cues listeners focus on when they judge the rhythmicity of metronome-based speech. As for the latter, Churaňová concludes that variation in the temporal domain of metronome-synchronized speech is more important for the perception of rhythm than the exact duration of rhythmic intervals. The results are of interest for speech applications insofar as they help to focus the modelling of rhythm on those phonetic parameters that actually play a role for listeners.

The chapter of Niebuhr critically scrutinizes the so-called "calling contours" at the end of utterances. These stepped intonation contours are characterized by two Fo plateaux: The first plateau spans the nuclear-accent syllable and leads over to a second lower plateau that extends until the phrase boundary. Based on a combination of speech production and perception experiments, Niebuhr presents evidence that the "calling contour" is not a single, phonetically and functionally homogenous category. Rather, there are three different subtypes of "calling contours".

They differ in the step up to the first plateau, as well as in the step down to the second plateau. Moreover, there are differences in the duration and intensity levels of the two plateaux, and even the degree of Fo flattening before the first plateau is not the same for the three subtypes. The communicative functions of the three subtypes of stepped intonation contours are outlined by Niebuhr with the terms "reluctance", "harmony", and "disharmony". In the long run, these prosodic insights will help improve automatic annotations and human-machine interactions.

Šturm deals with the complexity of articulatory movements in general and of the four Czech alveolar consonants [t,s,n,l] in particular. Based on electropalatographic (EPG) data collected in read sentences of nine native speakers, the results of Šturm's study reveal the huge amount of complexity and temporal and spatial variation in the production of alveolar consonants, ranging from differences between types of consonants to effects of the prosodic context. The fact that Šturm recorded nine speakers, which is far more than average in EPG studies, makes his results particularly reliable and additionally allowed him to provide rare, detailed insights into speaker-specific articulation patterns. Šturm's findings remind us that there is still a considerable gap between simple phonological models and organizations on the one hand and actual phonetic variation on the other. It is important for future phonological organizations and models to incorporate this variation, not least because it is often systematic and hence potentially functional in speech communication. Against this background, Šturm suggests ways to reduce the complexity of articulatory (EPG) data so that they become more handy to use in phonology and speech technology.

Landgraf's chapter is concerned with discovering the complexity of speech in noise, i.e. Lombard speech. While previous studies on this type of speech mainly focused on acoustic-prosodic measurements, Landgraf shows that increased noise levels also change the communication behaviour in general. For example, turns become longer, there are more hesitations, and the total number of turns in dialogues decreases as the noise level increases. Simultaneously, speakers become more active and produce more speech material under noise. Landgraf's study is embedded in a project on the linguistic evaluation of technical speech

enhancement systems for in-car communication. For this reason, she additionally compares the Lombard speech of an actual driving situation with the Lombard speech of a simulated driving situation. Her results show that there are only quantitative, but no qualitative differences between the two situations. The corresponding conclusion that real speech-production environments can be simulated in sophisticated laboratory settings opens up whole new possibilities for both the discovery of speech complexity and the development of speech enhancement technologies.

The volume section on the organization of complexity in speech starts with two chapters on the interplay of segmental phonetics and phonology by Howson and Monahan and Howson and Komova. Howson and Monahan present an acoustic analysis of voiced and voiceless fricatives in Czech, produced in lists of isolated words by male and female native speakers. The authors found that voiced and voiceless fricatives differ in spectral patterns and transitions, and that these differences fit in well with studies on voicing contrasts in other languages. Projected onto articulatory movements, the acoustic findings suggest that voiced fricatives are realized with a smaller pre-constriction volume that allows for both voicing and frication. Voiceless fricatives show a retracted tongue, which mediates the airflow produced from the open glottis. While these findings are relevant for speech synthesis and other applications, phonological frameworks are still incapable of modelling the complex articulatory interactions associated with voicing distinctions in speech.

Howson and Komova conducted a detailed ultrasound analysis of /r/ phonemes of Czech and Russian, realized at the onset, in the middle, and at the offset of target words. Measurements of tongue shape and tongue movement dynamics show that the plain /r/ is produced similarly in Czech and Russian, whereas the trilled Czech /r̝/ differs considerably from the palatalized Russian /rʲ/. The latter is characterized, amongst other things, by a more fronted tongue dorsum and a raised tongue body, which is indicative of an apico-laminal articulation. Moreover, unlike claimed in the literature, the empirical evidence of Howson and Komova suggests that the palatalization of /rʲ/ takes priority over the surrounding vowel context. The raised tongue body of /rʲ/ in combination with its apico-laminal articulation allow the assumption that the trilled Czech

/ɻ/ has evolved from the Russian-like /rʲ/ during a sound change. In this way, the study demonstrates how a better understanding of articulatory processes can contribute to explaining sound changes and defining the stability/clarity of sound categories in speech perception, which, in turn, has implications for speech technology.

The chapter of Valenta and Šmídl deals with confusions of words in the "Toll-Free Calls" corpus of Czech. The authors compare the word-identification performance of human transcribers to that of automatic speech recognition (ASR) systems. The results show – as expected – that the ASR performance is worse than the word-identification performance of human transcribers. However, while ASR systems process the spontaneous speech material in real time and a single pass, human transcribers worked about eight times slower than real time and also had the opportunity to play back the recordings repeatedly. Yet, humans also made mistakes so that inter-transcriber agreement hardly exceeded 90%. Relating the errors of humans to those of machines leads to a more realistic evaluation of the performance of ASR systems and additionally provides interesting implications for phonological models and cognitive speech processing.

Similar to Valenta and Šmídl, Volín and Bartůňková also address the performance of automatic speech processing systems. However, they look at prosodic rather than segmental aspects. More specifically, they evaluate the information value of simple (i.e., holistic) descriptors of F0 tracks. The study, embedded in the cross-linguistic question, examines the degree to which intonational differences between Czech and British English are contained in the speech of Czech L2 learners of English, and how well different F0 descriptors are able to uncover these L2 patterns. Results of an acoustic analysis show that Czech and English intonation differ along many F0 descriptors, such as F0 range, F0 level, and F0 declination, and that Czech L2 learners of English are for many descriptors in between the two groups of native speakers. Some findings run counter to the stereotypical attributes associated with Czech and English intonation. These inconsistencies have implications for phonological modelling. As regards the evaluation part of the study, Volín and Bartůňková conclude that all F0 descriptors, although simple to measure and only able to capture ho-

listic aspects of intonation, are sensitive and informative enough to be used at various levels of foreign-language analysis and teaching.

The section on the modelling of complexity starts with a chapter of Skarnitzl, Vaňková and Bořil on the question if and how vowel formant extractions can be optimized. To that end, the authors compare the two established software tools Praat and Snack (implemented in WaveSurfer), and test on the basis of 250 vowel tokens from male and female Czech native speakers whether the default formant analysis settings really perform best in the two programs, whether one of the two programs outperforms the other in terms of its automatic formant extraction algorithm, and how these algorithms perform in relation to manual human measurements. The results of this "competition" lead Skarnitzl et al. to the conclusion that both analysis tools work well. However, Praat is better than its reputation and in fact superior to Snack when it comes to automatic formant extraction, particularly of the second formant F2. The authors recommend specific formant-analysis settings for male and female speakers, and in this way make an important contribution to automatic data collection projects, which can then inform future models of speech complexity.

The chapter of Jůzová and Tihelka takes the opposite perspective and addresses automatic speech synthesis rather than automatic speech analysis. The authors compare two different approaches to speech synthesis: a regular diphone-based text-to-speech synthesis (TTS), and a special TTS variant that is able to concatenate larger units of speech thanks to using a corpus which is specifically tailored to the task of the synthesis algorithm, for example, reading weather forecasts. The authors call this TTS variant limited-domain (LD) synthesis. A small perception experiment shows that the LD synthesis indeed results in a better speech quality than the regular TTS system. Jůzová and Tihelka sketch on this basis which directions future developments of speech synthesis algorithms can and should take in order to be more acceptable in everyday applications.

At the heart of Vít's chapter is also a perception experiment. However, instead of directly judging the performance of a TTS system, Vít instructed his listeners to detect deficient sound segments in stimuli of short synthesized sentences, and to mark these segments on a PC screen. The results of the experiment are used to build a classifier that emulates

the human ear and can hence predict and correct potential TTS problems automatically before they actually occur. The author outlines what such a classifier can look like, to what extent it can improve future TTS systems, and where the limits of this approach lie.

The chapter of Borský and Pollák rounds off the range of technical speech applications by shifting the focus from TTS to ASR, i.e. automatic speech recognition. The authors present and discuss theoretical and practical aspects of the recognition of digitally stored data, particularly of sound files that were compressed by lossy encoders. For example, the authors investigate to which extent popular compression formats like MP3 and GSM interfere with ASR systems, and how potential ASR problems can be overcome or at least reduced. In this way, the final chapter of Borský and Pollák links back to the initial chapters of the volume, which were oriented towards the collection and phonetic-phonological analyses of speech data.

In summary, our volume is designed to draw on the two vertical pillars of the door frame into speech communication research, and not least for this reason, we hope that the twelve chapters – individually and in combination – will be able to open this door into speech communication research a little wider for many readers. In any event, the chapters clearly set the direction for the future of speech sciences: In the (early) 20th century, we saw that speech communication research diverged more or less widely into a more data-oriented phonetic branch and a more abstract, theory-oriented phonological branch. Today, we see – also in the chapters of this volume – that both phonetics and phonology, i.e. data collection, analysis, and modelling, grow together again under one roof; and simultaneously, they become more and more intertwined with research in engineering and technology.

Our image of a city's street and transportation network illustrated that languages develop over time, and so do the speech sciences. It is no longer sufficient for research on speech communication to accumulate empirical knowledge and develop theoretical frameworks. The financial and structural problems of phonetic and linguistic institutions all over the world make this fact abundantly clear. In order to overcome these problems, we will have to start thinking about how we can *sell* our

knowledge. That is, how we can collaborate with engineers, economists, physicians, social scientists, and people from many other fields in such a way that our undoubtedly substantial insights into the mechanisms of speech communication eventually have a positive impact on everyday life. Nass and Brave (2007) complain, not without reason, that "Interfaces that talk and listen are populating computers, cars, call centres, and even home appliances and toys, but voice interfaces invariably frustrate rather than help". Every corner of our life is filled with speech. Speaking with each other is the most important means of social interaction. This makes speech one of the most important research subjects of all.

So, the potential is there, and modern technologies offer – more than ever before in the history of speech sciences – great chances to realize this potential. The chapters in this volume show many interdisciplinary starting points for this bold venture, which is essentially based on "tackling the complexity of speech".

In this spirit, we, the editors of this volume, would like to express our deepest appreciation to all our authors for their excellent, multifarious chapters and – not less important – their timely submissions and revisions of these chapters. Speaking of revisions, we would also like to thank our reviewers Kristýna Poesová, Evelin Graupe, and Stephanie Berger for their insightful and helpful comments and suggestions on all chapters of this volume. Finally, we are also very grateful to our publisher for the efficient and competent handling of the manuscript.

Sønderborg/Prague, February 2015
Oliver Niebuhr and Radek Skarnitzl

References

Chitoran, I. – Cohn, A. C. (2009), Complexity in phonetics and phonology: gradience, categoriality, and naturalness. In: Pellegrino, F. – Marsico, E. – Chitoran, I. – Coupé, C. (eds.), *Approaches to phonological complexity*, pp. 21–46, Berlin/New York: Mouton de Gruyter.

Drosdowski, G. (1996), *Der Duden: Geschichte und Aufgabe eines ungewöhnlichen Buches*, Mannheim: Duden.

Jakobson, R. (1973), *Main trends in the science of language*, New York: Harper & Row.

Kohler, K. J. (1995), Phonetics – A language science in its own right? In: *Proceedings of 13th ICPhS*, pp. 10–17.

Kohler, K. J. (2000), The future of phonetics, *Journal of the International Phonetic Association* 30, pp. 1–24.

Kügler, F. – Smolibicki, B. – Baumann, S. – Niebuhr, O. – Wagner, P. – Peters, J. – Schweitzer, K. – Jannedy, S. – Grice, M. – Braun, B. (2015), DIMA – Annotation guidelines for German intonation. In: *Proceedings of 18th ICPhS, Glasgow, Scotland.*

Laver, J. (2001), The nature of phonetics, *Journal of the International Phonetic Association* 30, pp. 31–36.

Nass, C. – Brave, S. (2007), *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, Cambridge: MIT press.

Niebuhr, O. – Michaud, A. (2015), Speech data acquisition – The underestimated challenge, *Kalipho* 3, pp. 1–42.

Ohala, J. J. (2004), Phonetics and phonology then, and then, and now. In: Quene, H. – van Heuven, V. (Eds.), *On speech and language: Studies for Sieb G. Nooteboom* (LOT Occasional Series 2), pp. 133–140, Utrecht: LOT.

Ohala, J. J. (2010), The relation between phonetics and phonology. In: Hardcastle, W. J. – Laver, J. – Gibbon, F. E. (eds.), *The Handbook of Phonetic Sciences*, pp. 652–677, Oxford: Wiley-Blackwell.

Pellegrino, F. – Marsico, E. – Chitoran, I. – Coupé, C. (2009), *Approaches to phonological complexity* (Phonology and phonetics, Vol. 16), Berlin/New York: Mouton de Gruyter.

Sweet, H. (1911), *Phonetics*. *Encyclopædia Britannica*, 11th ed., Cambridge: Cambridge University Press.

Wagner, P. – Origlia, A. – Avezani, C. – Christodoulides, G. – Cutugno, F. – D'Imperio, M. – Escudero, D. – Gili Fivela, B. – Lacheret, A. – Ludusan, B. – Moniz, H. – Ní Chasaide, A. – Niebuhr, O. – Rousier-Vercruyssen, L. – Simko, J. – Simon, A.-C. – Tesser, F. – Vainio, M. (2015), Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In: *Proceedings of 18th ICPhS*, Glasgow, Scotland.

Winchester, S. (2003), *The Meaning of Everything. The Story of the Oxford English Dictionary*, Oxford: Oxford University Press.

View publication stats