Current developments in phonetics

Dijkstra, J.; Pols, L.C.W.; van Son, R.J.J.H.

Published in:
Conference Proceedings 'From Sound to Sense: 50 + Years of Discoveries in Speech Communication

Citation for published version (APA):
CURRENT DEVELOPMENTS IN PHONETICS

Louis C.W. Pols
Institute of Phonetic Sciences / ACLC
University of Amsterdam, The Netherlands
Louis.Pols@uva.nl

ABSTRACT
This contribution presents current trends in phonetics research with some bias towards the research of my own Amsterdam group. I also tried to exclude areas that will most probably be covered by other invited speakers. This let me to emphasize, apart from my admiration for Ken Stevens and his work, the importance of dynamic information in speech, not just in terms of formant transitions but also in terms of reduction, coarticulation, local context and information content. I also want to advocate the concept of computational phonetics, which allows an efficient type of modeling of speech events. Nowadays we can gain a lot of knowledge from annotated, and preferably freely available, speech corpora, of which a few examples are given. I am also proud of the role that Phonetics plays as an interdisciplinary science, most notable in speech and language technology, but also in signal processing, psycholinguistics, early speech development, multi-modal dialogs, and speech pathology, to mention just a few.

PHONETICS IN GENERAL
Phonetics is not just great because we have giants like Ken Stevens in our midst, but also because Phonetics is a fantastic interdisciplinary discipline, that confronts us with many tantalizing scientific questions, but also allows us to participate in and contribute to many aspects of life. Phonetics is about the speech signal, the spoken language and speech communication, it is about phonemes and prosody, it is about speaking and listening and about mental storage and retrieval, it is about speech acquisition and speech pathology, it is about speech technology and speech databases, it is about languages of the world and dialects, and it is about everything else that phoneticians are good at, like Laboratory Phonology, evaluating cochlear implants, or designing Web avatars. Because Phonetics is so highly interdisciplinary, phoneticians work in many different fields like Linguistics, Psychology, Pedagogy, History, Acoustics, Signal Processing, Artificial Intelligence, Computer Science, Medicine, Cognition, and probably several more.

From Sound to Sense: June 11 – June 13, 2004 at MIT
The phonetics community is a lively community, as can for instance be deduced from the 1000+ participants to international conferences like Eurospeech and the International Conference on Spoken Language Processing (ICSLP) (now jointly called Interspeech conferences and organized under the auspices of ISCA), the International Conference on Acoustics, Speech & Signal Processing (ICASSP), the International Conference on Language Resources and Evaluation (LREC), and the International Congress of Phonetic Sciences (ICPhS). But also from the numerous workshops that are organized every year on a great variety of topics, see for more details for instance the monthly ISCA Newsletter (http://www.isca-speech.org/), the monthly FoNETiks newsletter (http://www.jiscmail.ac.uk/lists/fonetiks.html) and the News section of the Speech Communication journal. Also the existence of the Linguistic Data Consortium (LDC) and the European Language Resources Association (ELRA) are healthy signs of our community. Another resource for activities in the field of Phonetics are the specialized journals, like Speech Communication, Journal of Phonetics, Phonetica, and Journal of the International Phonetic Association (JIPA), plus journals like The Journal of the Acoustical Society of America (JASA) and Journal of Language, Speech and Hearing Research (JLSHR), that always have substantial speech sections.

According to the written report of the most recent General Meeting in Barcelona (JIPA 33(2), 2003, 275-277), the International Phonetic Association (IPA) had, by the end of 2003, slightly more than 1000 members. Also the International Speech Communication Association (ISCA) presently has about 1350 members. I guess that the number of speech scientists all over the world is at least ten times larger. This does not yet make it a big community, but still a rather influential one. The number of serious books published in Phonetics and related fields must be of the order of 25 yearly. At the most recent 15th International Congress of Phonetic Sciences in August 2003 in Barcelona, Spain, where Ken Stevens was again one of the invited plenary lecturers, some 50 countries were represented, of which the largest number of (first-author) contributions came from the USA (158), followed by France (81), Germany (73) and the UK (71). The list of contributions from there until 12 continues with Japan (46), Spain (45), Sweden (41), Netherlands (31), Canada (25), Russia (19), Italy (17), Finland (14), Australia (12), Brazil (12), and China (12).

I am less optimistic about the number of university students that nowadays choose the Phonetics specialization. With the introduction of the Anglo-Saxon Bachelor-Master system we see, at least in Holland but this is probably symptomatic for many other countries, that students follow fairly general Ba-directions like General Linguistics, and that only a few specialize during the Ma-phase in Phonetics, Speech Communication, Speech Technology, Speech Development and Pathology, etc..
At the opening of the Eurospeech’95 conference in Madrid I had the honor, as then ESCA president, to present the ESCA medal for scientific achievements to Ken Stevens. In my search for interesting facts about Ken, I learned then that he had a record of one publication on average in JASA every year for over 40 years, apart of course from all his other journal publications such as the 1989 special issue of Journal of Phonetics devoted to his Quantal theory. The work of Ken Stevens is most prominently represented by his 1998 book on “Acoustic Phonetics”. I had the privilege to contribute the following recommendation to that book:

“A lifetime of knowledge and experience is bundled in this impressive book by the ESCA-1995 medalist Ken Stevens. All readers of this book will finally have the opportunity to share his detailed description of the production, the acoustics, and the perception of speech segments”.

His segmental approach is basic to all fields of Phonetics, but despite all his pioneering work there is still much more research going on in Phonetics and Speech Communication.

The organizers of this conference asked me to cover in my presentation the current developments in Phonetics. From the previous section you will understand that, in my perception, this would be a very wide topic to cover for which I would have to make specific choices. However, similar invitations went to other colleagues and they were asked to cover fields like ‘Phonology’, ‘Acoustics of Speech’, ‘Speech Planning and Production’, ‘Speech Perception’, ‘Development, Disorders and Remediation’, and ‘Technology: Synthesis and Recognition’. This suddenly leaves much less to cover under the specific heading of Phonetics! Still I managed to select a couple of topics for which I do have a strong personal interest, that most probably won’t interfere too much with the other invited speakers, and that still have the warm attention of many phoneticians, given their coverage at specialized workshops and at sessions at ICPhS. They are:

- Phonetics as a basic science
- Computational modeling, Computational Phonetics
- Knowledge from annotated, and preferably freely accessible, speech corpora
- Phonetics as an interdisciplinary science

From Sound to Sense: June 11 – June 13, 2004 at MIT
PHONETICS AS A BASIC SCIENCE

Phonetics in a narrower sense is often connected to phonemes and the like. The Summer Course in English Phonetics that is presented every year by the Department of Phonetics and Linguistics at University College London, for instance summarizes ‘all the main aspects of English phonetics’ in the following way:

- phonemic systems (vowels and consonants)
- segmental analysis (allophonic processes)
- word stress
- weakening and coarticulation processes
- sentence stress (accent, tonal stress)
- intonation and meaning

Most textbooks on Phonetics (e.g., Clark & Yallop, 1995; Hardcastle & Laver, 1997; Ladefoged, 2001a, 2001b; Lieberman & Blumstein, 1988; Pickett, 1999; Rietveld & van Heuven, 1997) somehow also use this type of subdivision.

While gathering at MIT, Cambridge again for this conference, I cannot help remembering that other remarkable event that took place here more than 20 years ago in October 1983: the Symposium on ‘Invariance and variability of speech processes’ (Perkell & Klatt, 1986). In October 1982 I had been appointed as a full professor in Phonetic Sciences at the University of Amsterdam, and I took this topic of ‘invariance and variability’ also as the Leitmotiv of our research for a number of years. Actually, I believe that it is still highly valid, and although we made some progress, many questions still remain. We made some progress on the perception of dynamic speechlike sounds, such as tone and formant transitions. Although the jnd for a stationary 1000 Hz tone is very low (about 1.5 Hz), the difference limen in the endpoint frequency of a short formant transition appears to be as high as 200 Hz (Van Wieringen, 1995; Van Wieringen & Pols, 1998). Van Son continued this work on speech dynamics and concluded that information in formant dynamics is only properly interpreted when the vowels are heard in an appropriate context (Pols & van Son, 1993; van Son & Pols, 1999). Van Bergem (1995) did interesting work on the perception of acoustic vowel reduction, such as in the Dutch word ‘miljoen (Eng. million)’ that can be pronounced as /mljun/ or as /m@ljun/. Subjects are very well
capable to specify whether the vowel /I/ is pronounced as a full vowel or as a schwa. Acoustic analysis of these stimuli indicated that vowel reduction appears to be much more contextual assimilation than centralization. Van Son did an interesting study on the efficiency of speech, by comparing, on the one hand, speaking effort (or reduction of that) in terms of segment duration, formant contrast and center of gravity with, on the other hand, the importance of the speech segment, expressed in terms of segmental information content in bits (Van Son & Pols, 2003; Van Son et al., 2004). It appears that more important (or less redundant) phonemes are less reduced, thus the organization of reduction according to conventional information structure does indeed increase speech efficiency. Also the importance of a structural factor like prominence (Streefkerk, 2003) becomes apparent in such an analysis.

An interesting paper that for me united past and present research was that of Kiefte & Kluender (2003) where they showed that results originally obtained with the Haskins Pattern Playback system about prevocalic voiceless stops consonant identification in CV syllables (Liberman et al., 1952), could not fully be replicated until a third-harmonic distortion, characteristic for the device used then, was added! I remember Peter Ladefoged indicating, while being in an animated discussion in front of that poster, that he was waiting for his, or other people’s, older results to be similarly challenged.

As my teacher Plomp rightly indicates in his fascinating book on the intelligent ear (Plomp, 2002), there are several biases that have affected the history of (speech and) hearing research: a dominance for simple signals (like sinusoidal tones) as stimuli; a preference for the microscopic approach (e.g., phoneme discrimination rather than intelligibility); an emphasis on psychophysical (rather than cognitive) aspects of hearing; and the use of clean stimuli in the laboratory, rather than the acoustic reality of the outside world with its disruptive sounds. It is reassuring to see that nowadays much more phonetic research is done on conversational speech and on the communicative and cognitive function of speech. It is also reassuring to see that people don’t give up to try to use specific phonetic knowledge to improve speech technology despite the overwhelming probabilistic approach preferred in that field. Examples are the use of specific durational knowledge in ASR (Pols et al., 1996), the use of prominence in TTS, and taking into account pronunciation variation, possibly by using an episodic model (Strik, 2003).

**COMPUTATIONAL MODELING, COMPUTATIONAL PHONETICS**

It is my impression that the requirements for properly performing speech technology have contributed substantially to the upsurge of computational modeling of speech events. An early example is the durational modeling that is required for speech synthesis (e.g., Klatt,1987; van From Sound to Sense: June 11 – June 13, 2004 at MIT
Santen, 1997). At the same time, duration modeling is already a rather complex example, because there are many aspects that control phoneme duration (such as phoneme identity, local context, position in the word, position of the word in the sentence, word stress, sentence accent, speaking style, speaking rate, emotion). Furthermore, segmental duration is only one of the aspects of prosody that have to be dealt with in speech synthesis, which also explains the existence of a book like ‘Computational Prosody’ (Sagisaka et al., 1997). The fact that so far very few speech recognition systems make proper use of prosody, is again an indication of its complexity. The more human-like dialog systems will become, the more important it will be to make proper use of prosody.

Figure 1: These stylized contours, taken from real utterances, show that the schwa in Dutch is not just a centralized vowel but is something that is completely assimilated with its phonemic context (adapted from van Bergem, 1995).

Roger Moore (1995) was one of the first to draw attention for computational phonetics. There are other interesting examples of systematically analyzing specific phenomena in speech production, which then can be properly modeled and then also allow for generalizations and predictions. Let me mention vowel reduction and schwa realization, as studied by van Bergem (1993, 1994, 1995b) for Dutch, see also Figure 1, and speech intelligibility as modeled by the Speech Transmission Index STI (Steeneken, 1992) A few other examples will be discussed in the next sections.
KNOWLEDGE FROM ANNOTATED SPEECH CORPORA

In the old days the only true knowledge was the one casted into rules. These rules were based on detailed measurements of carefully collected sets of speech material and/or on intuitions from the scientists involved. Nowadays it is more and more common to accept that specific knowledge can be derived from large databases and/or from an intelligent search in carefully selected subsets (e.g., greedy algorithm). This is for instance the way Van Santen (1997) determined his duration rules for speech synthesis. Greenberg and his colleagues (Greenberg, et al., 2003) use the Switchboard corpus to describe the peculiarities of truly conversational speech (over the telephone). In the Netherlands we are happy to have just completed the 10 million words Corpus of Spoken Dutch (CGN) (Oostdijk et al., 2002) as well as the IFA-corpus (Van Son et al., 2001) with fewer speakers (5 male and 5 female), but more and more diverse material per speaker. Within the Intas 915 project (de Silva et al., 2003; van Son et al., 2004) we compare phonetic properties of 3 widely different languages (Dutch, Finnish, and Russian), see Figure 2.

Figure 2: Correlation coefficients between global $-\log_2$(word frequency) and acoustic reduction (in terms of vowel duration, F12 distance, Center of Gravity, and intensity) for read speech (to the left) and for spontaneous speech (to the right).
PHONETICS AS AN INTERDISCIPLINARY SCIENCE

Young children easily accept that all attention of caretakers, family or friends is centered around them. Similarly there are nice postcards indicating that your university, or the small village you are living in or spending your vacation at, is actually the center of the world. Likewise do I have the feeling that many developments in science only could have happened because of the substantial influence that Phonetics and phoneticians had on it. Although this is certainly somewhat pathetic, I will give you some examples that may make you feel milder concerning this egocentric view of mine. The spectro-temporal representation of acoustic signals in the form of spectrograms, pitch extraction algorithms, formant extraction algorithms such as chirp-z-transform, FFT, LPC, ARMA, PSOLA, signal processing software (such as ILS, Entropics Waves, and praat), other software packages such as HTK, the CSLU speech toolkit, or the Mississippi State ISIP public domain speech recognizer. I also think of perceptual evaluative tests, such as the Diagnostic or Modified Rhyme Test, Mean Opinion Score, the use of semantically unpredictable sentences, paired comparison, absolute identification, discrimination and detection tests. Several of these tests came from psychology, or were developed for testing speech synthesis. However, they are just as useful to test cochlear implant patients, or laryngectomized people, or cleft-palate speakers, or early speech development. Also statistical and pattern recognition procedure have been improved and became more popular through speech science: principal component and discriminant analysis, canonical matching, template matching, vector quantization, Dynamic Time Warping, N-best recognition, phonetic decision trees, post-processing, HMM and probabilistic connectionist models for phoneme and word recognition, Adaptive Resonance Theory, dynamic programming and beam search, A*, n-gram language models, the concept of perplexity, and probably many more.

I will illustrate the interdisciplinarity of Phonetics with a few examples. One is the use of the source-filter model to describe early speech development in the first two years of life (Koopmans-van Beinum, 1990). This sensori-motor approach describes gestures of the infant speech production system rather than using the phonetic-linguistic approach in which phonemes are described. Gestures concern larynx movements (phonation) and supraglottal vocal tract movements (articulation). An utterance is defined as an infant sound produced during one breath unit. There are 5 phonation types (from ‘no phonation’ to ‘variation in intensity, duration, and intonation with interrupted phonation’) and 3 main articulation types (from ‘no articulation’ to ‘two or more articulatory movements in a two- or more-syllabic utterance’). Furthermore, for all utterances with one or more articulatory movements, a classification of place (front, central, back) and manner (fricatives and trills, stops including affricates, glides, nasals, and laterals) of articulation appeared to be worth while (Koopmans-van Beinum, 1999).
Vocalic sounds can be further differentiated in 4 degrees of jaw opening (close, close-mid, open-mid, open) and 3 positions of the tongue body (front, central, back). Movements in vocalic parts can be described, for instance from open/back to open/central. All of this also makes it easier to go from this level of differentiation to the more traditional phonological features for utterances of children in their second year of life when variegated babbles and recognizable words emerge.

Another example concerns the study of the production and evaluation of laryngectomized speech. In the Netherlands alone, around 700 individuals are diagnosed yearly with laryngeal cancer. When the tumor is in an advanced stage, a total laryngectomy is often necessary. In order to overcome a total loss of the natural voice, most of the time a voice prosthesis is placed in the fistula (created between the trachea and the esophagus). This acts as a one-way valve which enables the patients to use tracheoesophageal speech (TE). Part of the pharynx then serves as a new voice source called the neoglottis. Unfortunately, this neoglottis lacks much of the myoelastic-aerodynamic capabilities of a healthy glottis. Nevertheless some patients are able to produce a clear voiced/voiceless distinction and a varying pitch. Research is going on to perform acoustic analyses on TE speech, to use videofluoroscopy, digital high speed imaging and other methods to study the aerodynamics of TE speech. Also listening experiments will be performed to find out more about phoneme and word intelligibility, as well as about naturalness and prosody (Van As, 2001; Jongmans et al., 2003).

I will conclude with two more examples illustrating the interdisciplinarity of Phonetics, one concerns the role of turn-taking behavior in dialogs and the other the use of certain signal and data processing procedures to analyze objectively baby speech.

The shift in attention in Phonetics from isolated artificial stimuli to conversational speech, also implies more attention for various aspects of dialogs. An important linguistic and phonetic element then concerns the prediction of turn switches in conversation (Van Son, 2003). Van Son’s project studies the quantitative modeling of the identification of turn-relevant places (TRP’s) in conversation as an integration process of temporally unfolding information at different levels in speech, from conversation acts and semantics to prosody, phonetics, and visual cues. An interesting detail concerns the intended use of the laryngograph to detect the preparatory glottal closure that precedes most TRP’s.

It is a challenge to properly transcribe the progression in the vowel production of young babies. However, especially in the first year of life, their utterances are difficult to identify in terms of phoneme sequences, whereas also spectro-temporal analyses are difficult because of the
extremely high pitch of these voices. Reliable formant measurements are almost impossible and, if done nevertheless, generally are substantially biased by expectations. At our institute we developed a pitch-related bandfilter analysis (Wempe & Boersma, 2003) that runs automatically on unlabeled utterances and delivers up to 10 points per utterance in a (formant-related) plane derived from a principal-components analysis on all 40-dimensional bandfilter spectra. For more details see van der Stelt et al. (2003). Results for one of the five normal-hearing children and for one of the five severely hearing-impaired children are presented in Figure 3. Data sets at 5 months and at 24 months are displayed. For the normal-hearing child an expansion of the vowel space is clearly visible, whereas such a development for the hearing-impaired child does not take place.

Figure 3: Multiple spectral measurements within utterances of a normal-hearing child (to the left) and a hearing-impaired child (to the right) at 5 months (stars) and at 24 months (circles) in a principal-components reference space of normal-hearing children (background). One sigma ellipses are also indicated.

CONCLUSIONS
There are a few keywords that do characterize my contribution. One is the importance of dynamic information over stationary information, the other is the importance of local and global context for proper speech perception and understanding, a third is my pride in the
interdisciplinary nature of Phonetics that makes it a most valuable contributor to many scientific fields and to many society needs. Finally I believe that computational phonetics is a good way to model specific knowledge that will frequently be derived from annotated speech corpora.

REFERENCES


From Sound to Sense: June 11 – June 13, 2004 at MIT


