Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 24 (2001), 39-46.

ACQUIRING AND IMPLEMENTING PHONETIC KNOWLEDGE

Louis C.W. Pols

Abstract

Proper early acquisition of speech and language appears to be a necessary process to reach mature speech communication. In modeling the process of natural (and pathological) speech production and speech perception, we frequently concentrate on specific aspects of phonetic knowledge. But also to improve the performance of speech technological systems, an intelligent interpretation of the abundant, but sometimes also incomplete or absent, phonetic information is highly advisable. The above framework was used to discuss the progress made in speech science and technology over the past 30 years. This slightly modified paper [21] was originally presented as a plenary keynote at Eurospeech 2001 Scandinavia at the occasion of receiving the ISCA medal for scientific achievement.

1 Introduction

Acquiring and implementing phonetic knowledge, in my opinion, is a unifying concept in many fields of speech research. Babies who make their first steps in acquiring their mother tongue, just as second language learners, are confronted with it. Any psycholinguist, psychoacoustician, or phonetician studying normal and pathological speech production and speech perception concentrates on certain signal attributes and will try to model the consequences for speech perception. Speech technologists generally first try to stay away from specific phonetic knowledge, but if they really have to gain the final few percents in performance, this might be the area where some profit lays. This brings me to the following subdivision of important fields:

early speech and language acquisition (L1 and L2)
speech production and perception (normal and pathological)
speech technology

As far as (phonetic) knowledge is concerned I think it is appropriate to distinguish:

rules, regularities, analogies, synonyms
knowledge stored in databases

Knowledge is scattered and not uniquely defined. The physical characteristics of, for instance, a prominent word or of a question intonation are not defined once and for all. Rather, certain features have to be present or can be traded against others.

Acquiring knowledge is a life-long process for humans, only a small part of this knowledge is already available at birth [5]. Some systems are adaptive, but most speech machines go through a one-time training cycle after which their 'knowledge' is fixed. Confronted with a slightly different new situation, the machine generally lacks the flexibility to adapt to this new situation and to interpret whatever (other) information might be available.

Also knowledge *implementation* takes place in many different forms. It is a way of defining 'learning' and long-term memory. It is also a way to formalize certain findings, such as a model for vowel reduction [3], or for predicting speech intelligibility [27], or in the form of a network that predicts prominence from acoustic signal characteristics [29].

2 Early speech and language acquisition

Although early speech and language acquisition is certainly not my primary field of expertise, I supervised several projects in this area. My colleagues Florien Koopmans-van Beinum and Jeannette van der Stelt developed an ingenious system to describe early sound productions in terms of source characteristics and articulatory movements, thus creating a natural continuum between early sound productions, babbling, and the production of meaningful (pseudo-)words [14]. Much phonetic knowledge is reflected in this approach.

I find it especially intriguing to see how results on certain phonetic tests, such as phoneme identification and same-different judgments, might be early indicators for dyslexia [23]. Also the recent development of early hearing screening with babies may have serious implications for the way we test and monitor the early speech development towards mature language processing. If, through objective tests such as brainstem audiometry BERA, but in the future perhaps also via new imaging techniques such as PET scan and functional MRI, serious hearing defects can be detected early, one should also provide the means to monitor, train and test the early speech and language development (early detection requires early intervention!). Hearing aids for very young hearing-impaired children will be beneficial for speech and language development. But for optimal adaptation one cannot rely on subjective testing only. Also in developing and adapting digital hearing aids, there might be a role for speech scientists. The same is true for cochlear implants. Also in the whole discussion about cochlear implants for very young babies we may have a responsibility.

3 Speech production and speech perception

At Eurospeech conferences and otherwise, there is not so much interest anymore in presenting basic phonetic data about speech analysis, speech production and speech perception. What I am thinking of are experiments and procedures like phoneme identification, similarity judgments, trading relations, bite block experiments, effects of local context, speaker normalization, gating experiments, shadowing, matching and the like. Of course the experimental procedure is only a means to study certain phenomena, of which the experimental results then hopefully enrich our vision on speech production and perception. The various Ph.D. projects that I supervised in this field (see sect. 7) almost all produced interesting and sometimes challenging results. Let me just mention an alternative model for acoustic and lexical vowel reduction [3],

a description of perceiving dynamic speechlike sounds [33] and an acoustic description of consonant reduction [26].

4 Speech technology

Speech technology is a fascinating field of research and of applications, that barely existed 30 years ago. From a scientific point of view I see it as the test bed for all our acquired knowledge of speech phenomena.

If we would be able to produce fully natural sounding synthetic speech from reading any text in the appropriate style, then we would have solved the text interpretation and speech generation problem. We would do even better if we knew how to optimize synthetic speech for noisy communication channels, reverberant conditions, and elder or non-native listeners.

If we would be able to fully understand every word spoken in any utterance from any speaker and would be able to properly interpret the meaning of that utterance in its context, then we would have solved the speaker adaptation, the word recognition and the speech understanding problem. Similarly speech-to-speech translation, topic spotting, and document retrieval, just as multi-lingual applications, are challenging tasks to our limited knowledge.

If we would be able to properly accommodate laryngectomized speakers with an artificial voice, hearing-impaired and deaf listeners with alternative stimulation via optimized digital hearing aids, cochlear implants or otherwise, and blind readers with natural speech output, then we would have given a true contribution to understanding the needs for the handicapped.

5 Speech databases

(Annotated) speech corpora are our latest knowledge source. Both in the field of speech recognition and speech synthesis there were good reasons for collecting such material. Only gradually many people in the speech and language community realized that more can be done with such corpora than just training context-independent or context-dependent phoneme models, or just concatenating carefully selected smaller units to bigger chunks of synthetic speech. They happen to be valuable resources for duration modeling [32], for pronunciation modeling, for intonation modeling, etc. The popularity of the TIMIT database for speech research is at least partly related to its design and its accessibility.

As two representative examples let me shortly describe the 10-millions words Spoken Dutch Corpus (CGN) [17] and the much smaller but fully phonemically segmented IFA-corpus [25].

In a jointly funded project by the Dutch and Flemish government, about 1,000 hours of contemporary standard Dutch, as spoken by adults in the Netherlands and Flanders, will be collected over a 5-years period that started in June 1998. The fourth of seven releases was delivered in October 2001. Each release contains its share of audio files as well as one or more CD-ROMs with annotations, this latter part is updated with each release. Future distribution via DVDs is considered. The Dutch Language Union (http://www. taalunie.org/) holds all the rights and the distribution will be done via ELRA (http://www.icp.grenet.fr/ELRA/), it is quite unique that the consent of all speakers and of any other parties involved will have been acquired. In order to serve as many needs as possible from a wide variety of speech and language researchers with various backgrounds, the corpus will contain both monologues and

dialogues, both high-quality speech as well as telephone speech, both spontaneous and read speech, etc. The entire corpus will be orthographically transcribed by hand using 'praat' (http://www.fon.hum.uva.nl/praat/), and will be automatically lemmatized and annotated with part-of-speech information. A selection of one million words will be phonetically transcribed and syntactically annotated. One quarter of that will also be prosodically annotated. The CGN-website is a valuable source of information for all interested parties (http://lands.let.kun.nl/cgn/home.htm).

In this phase of the project, there is still more emphasis on how to collect and annotate all the material, then on how to make proper use of this CGN-corpus. A mid-term evaluation of the progress of the project is recently performed by BAS in Münich with highly positive results.

In compiling the 50,000-words IFA-corpus we built on the experiences gained so far within the CGN-project, but we also introduced several of our own demands. Most notably, this corpus contains much more speech (half an hour on average per speaker), of a higher recording quality, from less Dutch speakers (4 male and 4 female). However, per speaker various speaking styles (informal story telling, reading, retelling a read story, reading lists of sentences, words and syllables from the narrative stories) are available for comparison. Furthermore, all speech is phonemically segmented and labeled. Initial phoneme labels and segment boundaries were introduced automatically (on the basis of the CELEX pronunciation lexicon and a phone-based HMM recognizer) and then hand corrected by trained labelers. All compiled data tables are fed into a PostgresSQL database for subsequent data mining. Some initial results about speaking rates and phoneme durations are already available [25], but more intricate questions can hopefully be solved using the powerful query language SQL. The corpus is freely available and accessible on line under the GNU General Public License (http://www.fon.hum.uva.nl/ IFAcorpus/).

6 Thirty years ago

In September 2001 when Eurospeech 2001 - Scandinavia took place in Aalborg, Denmark, it was almost exactly 30 years ago that I had my first experience of being actively involved in a major speech conference. This was the 7th International Congress on Acoustics (ICA) in August 1971 in Budapest, Hungary. This event to me seems to be an appropriate landmark for comparing progress in our field since that time (for all authors mentioned below, see papers in the Proceedings of the 7th Int. Congress on Acoustics).

Many of the younger people will find it hard to imagine that in 1971 there was no text-to-speech synthesis-by-rule yet, diphones did not yet exist, only formant analysis-resynthesis using for instance Liljencrants' and Fant's OVE III. Rabiner's digital hardware for speech synthesis, using a cascade of 2-pole digital filters, was able to generate sentences like 'we were away a year ago', 'may we all learn a yellow lion roar', and 'few thieves are never sent to the judge'. Only isolated word recognition was possible with vocabulary sizes of some 50 words using template matching. Sakoe had just introduced dynamic processing for time normalization, Atal presented his initial ideas about predictive coding. The probabilistic approach was yet unknown. Dreyfus-Graf advocated an artificial language that would simplify automatic recognition. Erman was the first to talk about telephone input and Neely about speech recognition in noise, but their word sets were still limited to 54 isolated words, carefully spoken by one male speaker, being Ken Stevens. I used a list of 50 Dutch words spoken by 5 males to produce a dimensional representation of bandfilter spectra for recognizing the more or less stationary phoneme-like segments.

Kozhevnikov presented a paper about the perception of amplitude modulated vowel-like stimuli. Chistovich talked about vowel discrimination and she also gave one of the plenary lectures in which she emphasized the importance of psychoacoustics for speech perception. Also then already Flanagan gave a keynote on focal points in speech communication research. Sundberg showed real time pitch extraction of folk music, whereas Matthews demonstrated music synthesis. As nowadays is very popular with all the ISCA Tutorial and Research Workshops and other satellite events surrounding Eurospeech, then also we already extended our presence in Hungary with a most enjoyable 3-days Speech Symposium in the city of Szeged, in the southern part of Hungary close to the Bugac Puszta.

Another early scientific meeting made a large impression on me for various reasons, this was the Symposium on 'Auditory Analysis and Perception of Speech' [10], organized by Gunnar Fant and Ludmilla Chistovich in August 1973 in, what then still was called, the city of Leningrad. It was special because of the place and of the period in which it took place. It was also special because of the topic and because of the people involved.

One last memorable event that I want to mention here is the Symposium on 'Invariance and Variability of Speech Processes' [18], that was organized by Joe Perkell and Dennis Klatt in Cambridge, Massachusetts in October 1983. This was one year after I was appointed as a full professor in Phonetic Sciences at the University of Amsterdam. The search for systematic variability has lead my research ever since.

Name	year	early acq.	prod./perc.	sp. techn.
Loes Klaassen-Don (UL)	1983		x	
Gerrit Bloothooft (VU)	1985		x	
Louis ten Bosch	1991		x	
Herman Steeneken	1992			x
Paul van Alphen	1992			x
Mirjam Tielen	1992		x	
Amos van Gelderen	1992		x	
Cecile Kuijpers	1993	x	x	
Rob van Son	1993		x	x
Jeannette van der Stelt	1993	x		
Dick van Bergem	1995		x	
Astrid van Wieringen	1995		x	
Henning Reetz	1996		x	x
Xue Wang	1997			x
Irma Verdonck-de Leeuw	1998		x	
Paul Boersma	1998		x	
Sylvie Mozziconacci (TUE)	1998			x
Kino Jansonius-Schultheiss	1999	x		
Monique van Donzel	1999		x	
Jan van Dijk	2001		x	
Ahmed Elgendy	2001		x	
Corina van As	2001		x	

7 Joint research with Ph.D. students

I was privileged to be able to supervise since 1982 some 25 Ph.D. theses of which most were successfully completed, which in Holland implies the production of a good-looking booklet that is distributed world-wide to colleague scientists. Of course this work is generally also reported about at conferences, workshops and in the open literature. One reason for choosing such a wide theme for my keynote, was the fact that I wanted to emphasize that my scientific career only has been possible through intensive cooperative work with our students in many different fields of speech science and technology. Out of respect to all their hard work I include in the table above a list of all completed theses [1-9, 11-13, 15-16, 22, 24, 27-28, 30-33]. I also indicate in that table to which of the three main fields that I distinguished before (early speech acquisition; production and perception of speech; speech technology), they contributed most.

Especially for the early speech acquisition projects I received much help from my colleague Florien Koopmans-van Beinum, whereas for several other projects I shared the supervision, f.i. with Reinier Plomp, Ino Flores-d'Arcais, Aditi Lahiri, Egbert de Boer, Anne Baker, or Adrian Houtsma.

8 Developments over the last 30 years

Of course there has been a tremendous progress over the last 30 years. The almost 700 papers that have been presented at the Eurospeech 2001 conference alone, may well represent a larger amount of publications in our field than everything published in the whole of 1971. I expect that the number of people presently active in the field of speech science and technology will be well over 10,000 world wide. Some 1,000 of them came together in Aalborg and became a member of ISCA, this certainly is a powerful task force to solve communicative problems in our society.

However, the joint phonetic knowledge in the minds and in the PCs of these people is apparently far from sufficient to solve today's communicative demands. Telephone communication is everywhere but the speech quality of GSMs is deplorable. Speech is considered to be the most natural form of communication, but a proper dialogue with a computer information system is still in an experimental phase. We are able to produce intelligible synthetic speech, but naturalness is at stake and speaker and speaking style characteristics cannot be properly controlled [19]. Automatic speech recognition has made tremendous progress over the last 30 years, but still much remains to be desired, such as greater robustness and quick adaptation. Speech and language technology could be used more in education, language learning and aids for the handicapped.

Also many 'smaller' questions remain to be solved in the domain of phonetic sciences.

- How do listeners normalize over speakers? Even young children can cope with this problem, but we still don't know how they do it.
- How do listeners handle speech variation? Because of coarticulation, prosodic variation, speaking style, etc., no two utterances are ever the same. How do we interpret this variation? Is there any random variation or is there always a cause for any specific source of variation?
- What is a realistic front end processor that efficiently extracts all the relevant information from the speech signal, also for high-pitched voices or noisy speech?

- What are the processes in acquiring our mother tongue and foreign languages? What are the implications of a speaking or hearing defect, what of hearing aids and cochlear implants?

I feel privileged to have been part of this lively speech community for over 30 years and I have high expectations of progress in the years to come, partly because I have the impression that phonetic knowledge nowadays is more accessible and can more easily be implemented in descriptive models (e.g, computational phonetics [20]) and in technological systems.

9 References

[1] Alphen, P. van (1992): *HMM-based continuous-speech recognition. Systematic evaluation of various system components*, Ph.D. thesis Univ. of Amsterdam: 216 pp.

[2] As, C. van (forthcoming): *Tracheoesophageal speech. A multidimensional assessment of voice quality*, Ph.D. thesis Univ. of Amsterdam: 209 pp.

[3] Bergem, D. R. van (1995): *Acoustic and lexical vowel reduction*, Ph.D. thesis Univ. of Amsterdam: 195 pp.

[4] Bloothooft, G. (1985): *Spectrum and timbre of sung vowels*, Ph.D. thesis Free Univ. of Amsterdam: 169 pp.

[5] Boersma, P. (1998): *Functional Phonology. Formalizing the interactions between articulatory and perceptual drives*, Ph.D. thesis Univ. of Amsterdam: 493 pp.

[6] Bosch, L. F. M. ten (1991): *On the structure of vowel systems. Aspects of an extended vowel model using effort and contrast*, Ph.D. thesis Univ. of Amsterdam: 190 pp.

[7] Dijk, J. van (2001): *Mechanical aspects of hearing*, Ph.D. thesis Univ. of Amsterdam: 211 pp.

[8] Donzel, M. E. van (1999), *Prosodic aspects of information structure in discourse*, Ph.D. thesis Univ. of Amsterdam: 194 pp.

[9] Elgendy, A. M. (2001): A*spects of pharyngeal coarticulation*, Ph.D. thesis Univ. of Amsterdam: 312 pp.

[10] Fant, G. & Tatham, M. A. A. (1975): *Auditory analysis and perception of speech*. London, Academic Press: 564 pp.

[11] Gelderen, A. J. S. van (1992): *De evaluatie van spreekvaardigheid in communicatieve situaties. Globale beoordeling en gedetailleerde analyse van spreekprestaties van 11- en 12-jarigen*, Ph.D. thesis Univ. of Amsterdam: 265 pp.

[12] Jansonius-Schultheiss, K. (1999): *Twee jaar spraak en taal bij schisis*, Ph.D. thesis Univ of Amsterdam: 277 pp.

[13] Klaassen-Don, L. E. O. (1983): *The influence of vowels on the perception of consonants*, Ph.D. thesis Univ. of Leiden: 153 pp.

[14] Koopmans-van Beinum, F. J. & van der Stelt, J. M. (1998): "Early speech development in children acquiring Dutch, mastering general basic elements". In: S. Gillis and A. de Houwer (Eds.), *The acquisition of Dutch*, Amsterdam/Philadelphia, John Benjamins: 101-162.

[15] Kuijpers, C. T. L. (1993): *Temporal coordination in speech development. A study on voicing contrast and assimilation*, Ph.D. thesis Univ. of Amsterdam: 165 pp.

[16] Mozziconacci, S. (1998): *Speech variability and emotion: Production and perception*, Ph.D. thesis Technological Univ. of Eindhoven: 210 pp.

[17] Oostdijk, N. (2000): "The Spoken Dutch Corpus. Overview and first evaluation", *Proc. LREC-2000*, Athens, Greece, Vol. 2: 887-894.

[18] Perkell, J. S. & Klatt, D. H. (Eds.) (1986): *Invariance and variability in speech processes*. Hillsdale, NJ, Lawrence Erlbaum Ass.: 604 pp.

[19] Pols, L. C. W. (1998): "Foreword", In: R. Sproat (Ed.), *Multilingual text-to-speech synthesis. The Bell Labs approach*, Dordrecht, Kluwer Academic Publishers: xxiii-xxiv.

[20] Pols, L. C. W. (1999): "Flexible, robust, and efficient human speech processing versus present-day speech technology", *Proc. ICPhS'99*, San Fransisco, CA., Vol. 1: 9-16.

[21] Pols (2001): "Acquiring and implementing phonetic knowledge", *Proc. Eurospeech'01*, Aalborg, Denmark, Vol. 1: K-3-K-6.

[22] Reetz, H. (1996): *Pitch perception in speech: A time domain approach. Implementation and evaluation*, Ph.D. thesis Univ. of Amsterdam: 236 pp.

[23] Schwippert, C. E., Koopmans-van Beinum, F. J. & van Leeuwen, T. H. (1999): "Phoneme boundary perception in relationship to developmental dyslexia", *Proc. ICPhS'99*, San Fransisco, CA, Vol. 2: 877-880.
[24] Son, R. J. J. H. van (1993): *Spectro-temporal features of vowel segments*, Ph.D. thesis Univ. of Amsterdam: 195 pp.
[25] Son, R. J. J. H. van, Binnenpoorte, D., van den Heuvel, H. & Pols, L. C. W. (2001): "The IFA corpus: a phonemically segmented Dutch 'open source' speech database". *Proc. Eurospeech'01*, Aalborg, Denmark, Vol. 3: 2051-2054.
[26] Son, R. J. J. H. van & Pols, L. C. W. (1999): "An acoustic description of consonant reduction", *Speech Communication* 28(2): 125-140.
[27] Steeneken, H. J. M. (1992): *On measuring and predicting speech intelligibility*, Ph.D. thesis Univ. of Amsterdam: 165 pp.
[28] Stelt, J. M. van der (1993): *Finally a word: A sensori-motor approach of the mother-infant system in its development towards speech*, Ph.D. thesis Univ. of Amsterdam: 226 pp.
[29] Streefkerk, B. M., Pols, L. C. W. & ten Bosch, L. F. M. (2001): "Up to what level can acoustical and textual features predict prominence", *Proc. Eurospeech'01*, Aalborg, Denmark, Vol. 2: 811-814.
[30] Tielen, M. T. J. (1992): *Male and female speech. An experimental study of sex-related voice and pronunciation characteristics*, Ph.D. thesis Univ. of Amsterdam: 180 pp.
[31] Verdonck-de Leeuw, I. M. (1998): *Voice characteristics following radiotherapy: the development of a protocol*, Ph.D. thesis Univ. of Amsterdam: 137 pp.
[32] Wang, X. (1997): *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. thesis Univ. of Amsterdam**:** 190 pp.
[33] Wieringen, A. van (1995): *Perceiving dynamic speechlike sounds. Psycho-acoustics and speech perception*, Ph.D. thesis Univ. of Amsterdam: 256 pp.