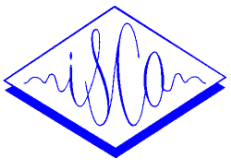

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Parametric High Definition (PHD) Speech Synthesis-by-Analysis: The Development of a Fundamentally New System Creating Connected Speech by Modifying Lexically-Represented Language Units

Hans G. Tillmann & Hartmut R. Pfitzinger

Department of Phonetics and Speech Communication
University of Munich, Schellingstr. 3, 80799 München, Germany
[tillmann|hpt]@phonetik.uni-muenchen.de

Abstract

Our paper has 5 sections. In section (1) we will discuss critically the fact that the development of Text-to-Speech systems and Speech-to-Text systems has in the past been treated as totally separate problems (we restrict ourselves to so-called dictation systems, L2S and S2L, which either translate written language units L into speech signals S, or speech signals S into sequences of written language units L). In section (2) we argue that for this reason, in the future, theoretical and empirical work should be devoted to providing an approach that integrates the L2S and S2L components into a unified phonetic system, which is able to learn to speak a language and also to understand what other L2S-systems are saying.

The new Munich PHD-system will be described in section (3) as an example of such a unified approach. Fundamental to this system is the selection and definition of lexically-given speech items, both acoustically and articulatorily (EMA). In section (4) we demonstrate a set of prosodic functions that take lexically-defined L-inputs and produce phonetically well-formed connected S-outputs. We discuss the possibility of combining certain elementary functions (such as those controlling F0 variation, segment duration, and sound modification) into a much more complex function which also controls the language-specific rhythmic variation of speech tempo in its locally measurable form. Finally section (5) will raise the question of analysing speech data produced by individual speakers as a means of arriving at the sound production system of a generalized representative member of the sociolect or dialect of the language in question.

1 Why Speech Technology has Treated Automatic Speech Recognition and Artificial Speech Synthesis as Two Separate Problems

The main reason for conceiving the Munich PHD Speech Synthesis-by-Analysis-Program as a foundationally new approach to phonetic speech research can be seen in the fact that in the past — with only few citable exceptions such as Bridle & Ralls 1985 [1], Hadersbeck 1988 [2] — speech technology has been treating the problem of relating speech signals (i.e. measurable data in the form of time functions) and speech categories (i.e. symbolic data in the form of printable characters) in two quite different versions. In the context of developing systems of automatic speech recognition we have the version where the problem takes the following form: what is given consists of measurable speech signals produced by the speakers of a language, what we ask for is the category of the perceived utterance in terms of the language of the speaker. For the development of speech-synthesis-systems the question is asked in quite the opposite direction: what is given is a categorical representation of an utterance of a language, and what is sought is a speech signal that will be perceived by the speakers/listeners of that language as fulfilling all the categories as defined by the printable text of that utterance.

As far as any text of any utterance can be related to a readable text in a given language we will use the term *text* as representing the category of an utterance. As far as any real speech utterance conveying the category of such a text must necessarily coincide with a measurable time function, we will use the term *speech* as representing the (digital) speech signal of a given utterance. Any regular speech utterance is thus a phonetic fact, consisting of a measurable speech signal and a related printable category.

From a purely technical point of view it really makes sense to handle speech recognition and speech synthesis as two separate problems, referred to as Speech-to-Text, S2T, and Text-to-Speech T2S. One reason surely is that S2T is much more a bottom-up problem than T2S is. On the other hand top-down-components play a much greater role in T2S than they do in S2T. In S2T we have to define (and than also to detect) necessary conditions for deciding which words of the language could have been produced in a given utterance (if it is a clear regular utterance of that language), whereas in T2S it will be often adequate enough only to define sufficient conditions for creating a speech signal that fulfils only some of the conditions needed by the listeners to perceive the required text category.

2 Two Main Goals and the Distinction Between Two Types of Speech Acts

Our decision to create a new research system at the University of Munich was motivated quite programmatically. The first programmatic component of this decision was to combine two different goals, a more practically oriented and a more theoretically oriented one. The second component consisted in introducing a clear distinction between two separate sets of speech acts.

The practically oriented initial main goal was given by the plan of systematically trying to apply the newest methods of speech technology to phonetic speech research, and to do this as effectively as possible in order to gain a better understanding of the facts that determine natural spoken language processing. The other main theoretically oriented goal was to use all available technologies of data and information processing for the development of so-called CPTs, i.e. complete phonetic theories of a spoken language, the idea of which was presented in Tillmann & Pompino-Marschall 1993 [13].

The second component mentioned above had to do with the fact that speech research has to deal with a variety of quite different kinds of speech acts. But there is a major distinction to draw between what could be called natural and non-natural speech acts, NSP and NNSP. Future speech technology will also have to consider this distinction.

Spontaneously produced normal speech is a typical case of NSP. One of the most relevant questions that has been asked by philosophers of language (Kemmerling 1980 [4]) deals with the remarkable observation that in NSP a speaker can express a great amount of information by uttering only a small amount of words: How is it possible that a speaker utters only a few words in order to make the listener understand states of affairs which need many more words in any explicit description?

But there is also quite another set of speech acts where what a speaker wants to express is nothing else but just the words of his utterance. A typical example of this type of NNSP is dictation. Another one can be found in teaching situations where the speaker wants to demonstrate the phonetic form of a word that the listener is asked to copy in his own utterance.

The distinction between NSP and NNSP is so relevant for speech research simply because the phonetic form of the respective utterances is so different. Citation forms, produced in NNSP, possess an alphabetically explicit structure showing the printable categories in a much clearer way than spontaneously produced strongly reduced forms of NSP do [12].

3 A Unified Approach to L2S and S2L

We have called PHD a synthesis-by-analysis-system because our aim was to model the speech-text-relation in both directions. For the time being we are restricting ourselves to the development of dictation systems, in both directions, from speech to text and from text to speech. The category of text is represented by lexical units L, produced by an individual speaker, whose speech production is investigated, i.e. analysed and then parametrically synthesized.

The speech of a speaker under investigation is given in two different forms, as an articulatory speech movement <ASM> and as a "soundstream". Articulatory speech movements (ASM) are characterized by angled brackets, soundstreams (SST) by double quotation marks. In the Web version of this paper available at `www.phonetik.uni-muenchen.de` these angled brackets and quotation marks also represent links enabling the reader to inspect the corresponding data. Soundstreams in quotation marks can be heard as an audio signal and inspected as a sonagram. Angled brackets refer to EMA data, which can be inspected as a partial representation of the underlying articulatory movements. In the framework of the PHD approach we consider soundstreams such as "bu:x" as the acoustic mapping of the underlying articulatory speech movement <bu:x>, even if the EMA data presented in this link contain only a partial representation of the complete processes in the production of the German word *Buch (book)*. To understand the differences in the phonetic forms of a word uttered either in isolation or in connected speech we should always look at the soundstreams and articulatory speech movements in relation to each other (see Tillmann 1998 [11]).

We also consider the citation form of phonemes, produced by a given speaker in single NNSP-syllables, as lexical units L. So a speaker of German may produce the consonantal soundstream "be:" or "ax", and the vocal soundstream "u:" in order to demonstrate the alphabetic components of the word *Buch*; see also the corresponding ASMs <be:>, <u:>, <ax>. Here, purely phonological questions can also be discussed. Speakers of German can only demonstrate the long tense vowels of their language in isolated forms as lexical units L. So "o:" is also understood by naive speakers of German as representing the lax (short form) "ɔ", which is also orthographically represented by the letter *o*. In the PHD framework the alternative is either to derive the short vowels as a prosodic modification of the long ones or to represent them directly by short VC soundstreams such as "ɔb" with ASM <ɔb>, which is the realization of the word *ob (whether)*.

Our synthesis-by-analysis speech research program is to use the new PHD-systems in a straightforward way to answer questions which have been asked for more than 100 years. There is no better characterization of the focus of our interest than the title of Rousselot's PHD-dissertation [9, 10]: *Les modifications phonétiques du langage*, i.e. the regular variability of the phonetic form of selected L-units. Rousselot looked at lexical items L of French. We are interested in how the SSTs "l", "o:", "ax" with the related ASMs <l>, <o:> and <ax> have to be modified to deliver the proper soundstream "lɔx", the German word *Loch (hole)* containing the acoustic picture of <lɔx>. Quite another question is to find the parameters that have to be modified in a proper way to put a word like "und" in soundstreams like "Maluma und Takete", "ja und nein", "neunundneunzig" with the corresponding speech movements < >, < > and < >.

4 Prosodic Speech Functions

In our synthesis-by-analysis PHD-system we ask the classical question of how lexical entities change their phonetic form under given prosodic conditions. To answer this question we use SLP-tools not only to systematically compare the resulting soundstreams and articulatory speech movements.

The second author has also developed a complex set of DSP-programs that allow the soundstreams of isolated L-units to be morphed into the reduced forms of connected speech. So the SST of "und" (with the ASM <und>) is parametrically transformed into the soundstreams "ja und nein", not only in normal speech,

but also in fast and slow speech: Listen to the slow versions “ ”, “ ”, the fast versions “ ”, “ ”, and look also at $< >$, $< >$ and $< >$, $< >$.

The set of DSP-programs which will be described by the second author in more detail in his forthcoming PHD-thesis not only allow us to answer the new question of what the regular modifications look like that are needed for integrating single lexical items L in the soundstream of connected speech, but also enable us to give the answer by explicitly defining the values of the parameters that control the observed modifications. Thus we can develop an algorithm in terms of a prosodic function *prof* that takes L-units such as “Maluma”, “and”, “Takete”, “9”, and “90” as lexical inputs and transforms them according to an abstract prosodic contour, into a sequence of properly connected words (see fig. 1). Thus, given a sequence of words as produced by a given speaker and an adequately generalized prosodic function, *prof*, which properly correlates the parametric F0-, intensity- and quality-reducing-values with the local speech rate of the desired utterance, the PHD-system will compute the proper soundstreams of “Maluma und Takete” or “99”, also in fast (“ ”, “ ”) and slow speech (“ ”, “ ”) [6, 7, 8].

5 From Individual Speakers to a Generalized System Controlling Interindividual Variation

If a soundstream SST can be computed as a function

$$\text{“SST”} = \text{prof}(\text{pros}_j, \{L_i\}), \quad i = 1, \ldots, M \text{ and } j = 1, \ldots, N,$$

where pros_j is a selected prosody and $\{L_i\}$ is a sequence of words, then we could try to generalize over a sufficiently large set of individual speakers of a given sociolect of a language in such a way, that every individual speaker γ can be modelled by a meta-function Γ,

$$\gamma_n(\text{“SST”}) = \Gamma(\gamma_n, \text{prof}(\text{pros}_j, \{L_i\}))$$

which then produces any desired soundstream “SST” in the voice of one of the individual speakers that the system has learned to map.

At present, we use the PHD-system for dictation of non-natural speech acts to audiences in order to expand our phonetic knowledge which, in turn, is expected to improve automatic speech recognition.

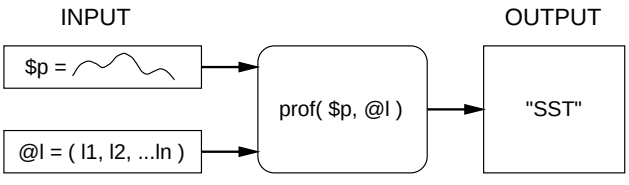

Figure 1: Scheme of the production system of soundstreams.

REFERENCES

[1] Bridle, J. S.; Ralls, M. P. (1985). An approach to speech recognition using synthesis-by-rule. In Fallside, F.; Woods, W. A., eds., *Computer Speech Processing*, pp. 277–292. Prentice Hall, Englewood Cliffs.

[2] Hadersbeck, M. (1988). Die Entwicklung eines Synchronisierenden Artikulators als wissensbasiertes Computerprogrammsystem PHONEX. Forschungsberichte (FIPKM) 26, Institut für Phonetik und Sprachliche Kommunikation der Universität München.

[3] Jones, D. (1931). The “word” as a phonetic entity. *Maître Phonétique*, 46(36): 60–65.

[4] Kemmerling, A. (1980). How many things must a speaker intend, before he is said to have meant? *Erkenntnis*, 15: 333–341.

[5] Kohler, K. J. (1998). The disappearance of words in connected speech. In *ZAS Papers in Linguistics*, vol. 11, pp. 21–33, Berlin.

[6] Pfitzinger, H. R. (1999). Local speech rate perception in German speech. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 2, pp. 893–896, San Francisco.

[7] Pompino-Marschall, B.; Piroth, H.-G.; Hoole, P.; Tillmann, H. G. (1984). ‘Koartikulation’ und ‘Steuerung’ in der Wahrnehmung des ‘momentanen Tempos’. Forschungsberichte (FIPKM) 19, pp. 306–314, Institut für Phonetik und Sprachliche Kommunikation der Universität München.

[8] Pompino-Marschall, B.; Piroth, H.-G.; Tilk, K.; Hoole, P.; Tillmann, H. G. (1982). Does the closed syllable determine the perception of ‘momentary tempo’? *Phonetica*, 39: 358–367.

[9] Rousselot, P.-J. (1891). Les modifications phonétiques du langage. *Revue des patois gallo-romans*, 4: 65–208.

[10] Tillmann, H. G. (1994). Early modern phonetics, especially instrumental and experimental work. In Asher, R. E.; Simpson, J. M. Y., eds., *The Encyclopedia of Language and Linguistics*, vol. 6, pp. 3082–3095. Pergamon Press, Oxford.

[11] Tillmann, H. G. (1998). Why the word should become the central unit of phonetic speech research. In *ZAS Papers in Linguistics*, vol. 11, pp. 1–20, Berlin.

[12] Tillmann, H. G. (*forthcoming*). Phonetic facts. Foundational observations concerning utterances in natural and non-natural speech acts.

[13] Tillmann, H. G.; Pompino-Marschall, B. (1993). Theoretical principles concerning segmentation, labelling strategies and levels of categorical annotation for spoken language database systems. In *Proc. of EUROSPEECH ’93*, vol. 3, pp. 1691–1694, Technische Universität Berlin.