Applying the Munich Parametric High Definition (PHD) Speech Synthesis System to the Problem of Teaching Chinese Tones to L1-Speakers of German

Hans G. Tillmann [a] *& Hartmut R. Pfitzinger* [a,b]

[a] Institut für Phonetik und Sprachliche Kommunikation, University of Munich, Germany
[b] JST/CREST at ATR Human Information Science Laboratories, Kyoto, Japan
tillmann@phonetik.uni-muenchen.de; hpt@phonetik.uni-muenchen.de, hrpfitz@atr.jp

Abstract

The aim of this paper is threefold:

- First and foremost we would like to propose a new strategy of phonetic speech research that can be distinguished from more traditional research programs in more than one important aspect.
- As directly opposed to the classical analysis-by-synthesis paradigm our second aim is to install the new paradigm of synthesis-by-analysis.
- And our third aim is to convince our audience that only strictly application-oriented phonetic speech research will lead to a deeper understanding of how speech acts really function phonetically.

The paper will be given in two parts. After the presentation of the philosophy behind our new research activity mentioned in the title, the second author will present a demonstration of the PHD-system in order to illustrate its applicability for solving the problems described by the first author.

1. The challenge: A long term application-oriented phonetic research program

In spoken language processing (SLP) one of the most provoking challenges of speech-based man-machine communication is the development of intelligent automatic systems for teaching well-motivated individual learners to speak a new foreign language fluently after as short a training time as possible.

Inspired by the work of the pioneers in this new research field (such as Rodolfo Delmonte, Farzad Ehsani, Maxine Eskenazi or Stephanie Seneff, to mention only a few names in alphabetic order) and looking at already existing Spoken Language Learning Systems (such as Delmonte's [5] or the MIT SLLS for Mandarin [4]) and with respect to the fact, that in the well established research field of second language acquisition there is not yet much knowledge available for reducing foreign accents [6] — at least concerning very 'narrow' phonetic conditions — we have come to the decision to propose the following application-oriented phonetic research strategy which starts with the following steps: (i) use the existing technology of SLP and develop a laptop-system that teaches L1-speakers of German to reproduce the tones of Mandarin Chinese; (ii) begin with single syllables and isolated words to demonstrate their tones in a very clear form; (iii) take the reproductions of the learning speaker, analyze them and, if necessary, modify them into a corrected form; (iv) present the original form as well as the corrected one, both in the voice of the individual speaker, to the learner so that he immediately can compare both of them directly in order to see (on the screen) and hear (on his headphones) the relevant differences. In many repeated such sessions the brain of the learner has to develop the neural programs for recognizing and reproducing the relevant categories that are to be distinguished in the new language (cf. for instance [3] where L1-speakers of Japanese were trained to master the /r/–/l/-distinction in prevocalic syllable position).

But these, of course, are only the first steps towards teaching the pronunciation of Mandarin tones to the learner. The tones of Chinese change their form as soon as the lexical items are uttered in connected speech [14]. It is probably for this reason that up to now tones are not taken into account in the automatic speech recognition technology of this language. Here it becomes clear why the development of the proposed automatic system must be seen as a phonetic research tool. The correct forms are to be parametrically analyzed and properly transferred to the complex utterances in the voice of the individual learner. And in this way we get the data collections which are needed for developing the phonetic theory of proper speech production in incremental steps and — at the same time — to scale and enlarge the domain of the teaching material.

Application-oriented speech research of this kind has two happy side-effects.

1. The application itself can be described quite naturally in ordinary language (and everybody will then also be able to judge whether the application indeed works satisfactorily). This aspect is very helpful for getting the necessary funding. And there is a very convincing argument which says that anybody who has brought himself into the situation of being able to pronounce the words of a new foreign language correctly will also start to speak this language in a much shorter time.
2. But in order to be able to successfully approach such an easily describable application effectively, one has to translate the ordinary language description into the technical terms outlining the tough problems requiring resolution by means of scientific work. And here the second side-effect consists in the fact that the resulting set of problems that must be solved is so rich in its complex structure that nobody could invent it without the given application. This will be discussed in the following sections.

It should be mentioned that our phonetic research proposal has been motivated by two technical developments in Munich. The first one has to do with the experience we have gained with the Munich Automatic Segmentation System (MAUS) [1]. It allowed us to reliably segment and annotate spontaneous speech

in terms of the actually realized sound elements. And this works because speech recognition technology is only used for the purpose of speech verification (knowing the text of a given utterance the canonical form of the lexical items in the spoken utterance can be identified for deriving and generating all possible and also even impossible pronunciation variants; so the actually given phonetic facts will be verified). The second argument for our proposal can be seen in the fact that all the available SLP-technologies of digital speech signal processing can be now used in a controlled manner for proper complex modification of a parametrically analyzed naturally produced utterance [11].

2. Speech acts and utterances

Traditional speech scientists have pointed out the trivial fact that there is no natural speech act without the concrete utterance of a given speaker (but they were mainly interested in analyzing any given utterance with respect to the phoneme system of the language of the speaker: take Bloomfield's first definition of his Set of Postulates [2] for the science of language or read Trubetzkoy's first sentence in his Principles of Phonology [13]). Future speech scientists will have to look at the great variability of different speaking styles and will have to try to give a precise answer to the question of what kinds of speech acts can be inferred from the phonetic form of any given utterance if it is produced by a real speaker.

Given the context of our research proposal we only have to distinguish, in a first step, two quite different kinds of speech acts. This clearly depends on the intention of the speaker. In the first case the speaker wants to present to the audience (or to himself) nothing else but the utterance itself. In this very special case we logically get the autonymical form of an utterance. The meaning of such an utterance is true, if it demonstrates (or presents the instantiation of) a phonetically correct form of the given category. Clear speech in a dictating context can serve as a good example of autonymically produced utterances.

In normal speech situations the speaker automatically transcends the utterance he is producing in order to semantically master the concrete or abstract situation he is acting in. In this case the utterance is logically produced in a heteronymic form. The speaker produces in an act of speech an utterance in order to express himself without taking care of the phonetic form of his utterance. The distinction of an autonymic vs. a heteronymic use of speech utterances is central and crucial for our research proposal because autonymically and heteronymically used forms are phonetically pronounced quite differently.

On the other hand it should be also mentioned here that in all the well-known philosophical theories of speech acts utterances don't really play a central role. Philosophical speech act theorists try to answer the question why and how it is possible that a speaker — just by only uttering "p" — can effectively express the meaning of p or even can convey the truth of a proposition p. For them the production of an utterance "U" seems not to be any problem at all. In this situation we have to realize that the production of an autonymically or heteronymically used utterance is still an unsolved empirical problem that can never be explained seriously by only a philosophical theory.

3. The complexity of phonetic facts

The traditional term 'utterance' as used by linguists or philosophers is systematically ambiguous. It has at least two empirical meanings. This is because we have to take into account two different kinds of realities. First we must take the utterance — especially in its autonymic form — as what is (or can be) directly perceived by the speakers and the listeners during an act of speech production. On the other hand there is the physical world which remains trans-phenomenal to the speakers and listeners of any natural speech act. It is this second reality where the phonetic speech scientist can derive the time functions of the speech signal and store them in a digitized form on a disk. Perceived utterances cannot be stored in this manner. They are always bound to a perceiving subject, but for the perceiving subject they have a category which can either be demonstrated by another autonymic categorical reproduction or they must be named by the use of a symbolic representation.

In logical terms, the relation between these two kinds of empirical data — categories and time functions, symbols and signals — is in not an analytical one. Because they are logically independent of each other and given as contiguous facts, we may say that they are empirically related in a very strong fashion: The categories of a speech database can be experimentally reproduced by just reproducing the time function that belongs to the given speech utterance.

On this theoretical background we are now in a position to introduce the concept of phonetic facts. A phonetic fact is the utterance (of a speaker produced in a speech act) that has a certain category and an empirically verifiable time function. Both can be stored in a spoken language database as a categorically annotated digital speech signal.

From a cognitive point of view we can say that during any speech act the speaking nervous system produces at its periphery very complex speech movements that are — only partly visible — observed by the sensory systems of speakers and listeners in order to identify the complex category of the observable phonetic facts that establishes the given speech act, be it phonetically an autonymic or heteronymic one. Any act of speech is complex even if the speaker produces only a single speech sound or the citation form of a short word. So also the demonstration of a cardinal vowel produced with a certain tone of a certain tone language is by itself a very complex action.

If such a nearly elementary categorical autonymic action is integrated in an utterance of much greater complexity we get functional variability. Functional variability in a naturally produced CVC-stream is controlled by different prosodies which depend on the act of speech in a given semantic and pragmatic situation. We do not yet know much about the rules that determine the complex structure of functional variability of phonetic facts. (So again: Future speech scientists will have to find a precise answer to the question of what different kinds of speech acts can be inferred from the phonetic form of a given utterance!)

For automatic speech recognition the variability of the phonetic form of lexical items in fluent speech caused by prosodic variation (such as local speech tempo and the resulting rhythmic structure of a naturally produced utterance in a given speaking style) is still a severe unsolved problem. Trying to take all possible pronunciation variants into account would not only lead to an explosion of computing time, but it is also doomed to fail because it seems impossible to provide the necessary training data covering all possible prosodic pronunciation variants.

This is quite different in the case of our proposed new research program, because in well-defined teaching situations all necessary pronunciation variability can be provided for proper verification as well as for the controlled modifications for generating the prosodically corrected forms.

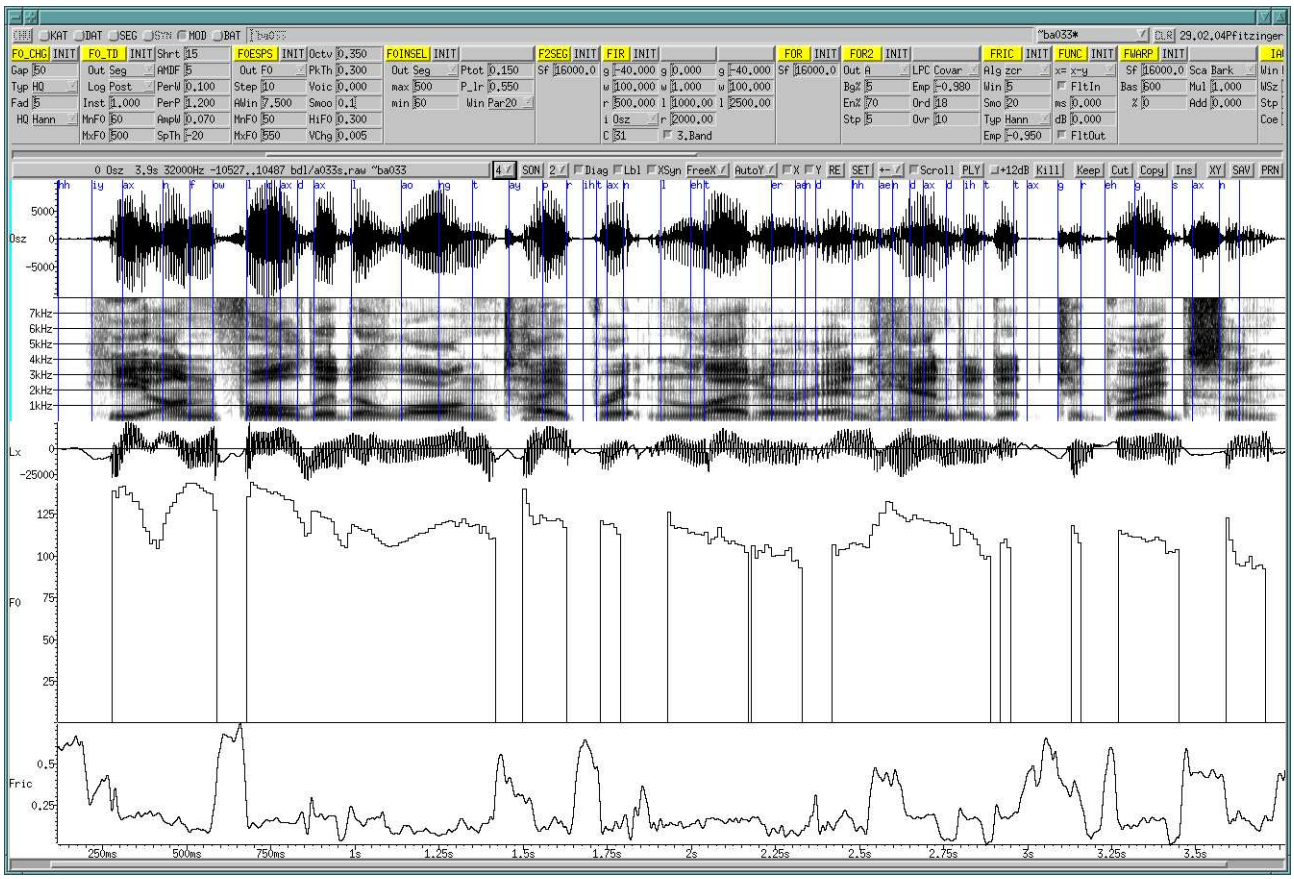

Figure 1: User interface of the PHD-system: In the top line the speech database, processing mode (concatenation, wildcard-search, segmentation, modification, batch), and selection criteria are determined. This Figure exemplifies the modification mode, which loads and displays all modification tools (which are executed by clicking on one of the yellow buttons). The remaining part of the user interface displays original signals as well as optional transformations for further interactive analysis, audio playback, or printing.

4. Speech synthesis by parametric analysis and controlled complex modification

The classical analysis-by-synthesis paradigm of phonetic speech research has been an excellent strategy be it for determining and characterizing elementary categories such as the voice onset times of voiced vs. unvoiced stop consonants or be it for discovering the relevant formant transitions of phonemes which differ in their place of articulation. On the other hand, the analysis-by-synthesis strategy had the severe consequence of totally reducing the categorical richness which is owned by any naturally spoken utterance, even if it is only the autonymic demonstration of a German sound sequence in the voice of the first author in a citation speaking style like “bedege” or “abetse”.

This drastic impoverishment of any categorical information bearing variability of any naturally produced utterance is probably one of the main reasons why also the classical source-filter-oriented parametric speech synthesis did not produce the desired results. The categorical abundance of naturally created phonetic facts cannot be easily simulated by any simple rule-based system for modeling the given control parameters. It is the complex prosodic picture as a whole that is owned by the complex and gives it its categorically determined character in a way that cannot be reproduced by an elementarily oriented parametric speech generation system.

We are deeply convinced that any parametric speech high quality artificial speech synthesis system has to start with complex utterances produced by real individual speakers. One possibility would be to take from a well prepared database the digital signal of some citation forms as a list of lexical items (or even only parts of them, or even only isolated canonical demonstrations of elementary speech sounds) in order to analyze them parametrically and then modify these alphabetically explicit phonetic forms so that the prosodic variant results that is functionally correct with respect to the intention that should be expressed in the given act of speech.

It is still a long way to go in phonetic speech research before we can begin to understand all the modifications that natural speakers generate as soon as they take a sequence of words out of the list of their mental lexicon and change their phonetic form in exactly the way they do in reality. We are deeply convinced that the new paradigm of speech synthesis by parametric analysis of complex phonetic facts (especially the locally variable speech rate [8]) in order to produce a prosodic modification of that utterance will give us a research strategy that can lead to the insights that are needed for future parametrically controlled speech generation systems. Even the categories of the speakers can thus be modified (an early example is [12]).

Modification of any given alphabetically explicit form into a prosodically correct reduced form or even an emphatic expanded form cannot be produced by linear interpolation or

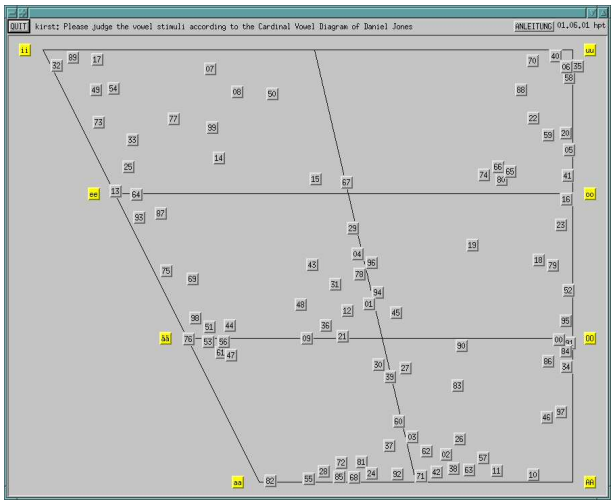

Figure 2: PHD-tool for interactive perception experiments: In this example the user interface was configured for conducting vowel identification tests. (These results were presented in [8].)

extrapolation. Believing in Paul Menzerath's early theory of speech production (according to which coarticulated speech movements are complexly organized in order to produce a sensory result on the articulatory and auditory retinas of speakers and listeners) we hope that the EMA-data we are collecting with the new 3D-system [15, 10] will help us to understand why the prosodic forms in the acoustic picture of the speech movements look as they look. We have already seen that a speaker can produce a high local speech rate either with slower and reduced articulator movements or with faster and more extensive ones. But this is audible as a categorically different speaking style.

5. Demonstrations of the PHD-system

Fig. 1 shows the user interface of the PHD-system. When set to 'modification' mode all Unix-shell-based signal processing tools are graphically displayed allowing fast and easy interactive access. All induced transformations arising from an original signal selected from any spoken language database are logged and converted to a batch file. This can then be interactively modified, improved, and finally applied to the entire speech database at the Unix-shell-level. The PHD-system can be regarded as a powerful higher-level speech analysis development toolkit. Also, in 1998 we implemented a freely configurable tool for conducting perception experiments (Fig. 2). Since then we used it in various studies [7, 8, 9] to get training data for the development of automatic feature detectors. More information as well as examples can be downloaded from: `http://www.phonetik.uni-muenchen.de/~hpt/`

6. Conclusions

This work proposed a new strategy of phonetic speech research which is significantly different from more traditional research programs. In the given state of the art of phonetic speech research we believe that the proposed research program of developing second language training systems can help to define a good framework for a large scale cooperation of phoneticians and SLP-specialists. This might even one day lead not only to a better quality of TTS-systems but could also be fruitful in producing results which are still needed for improving existing ASR technology.

7. Acknowledgements

The development of the PHD-system is partly supported by the BMW Group Research and Technology Pty Ltd, Munich in the project entitled "Principles for future improvements in speech synthesis in car environments".

8. References

[1] Beringer, N.; Schiel, F. 2000. The quality of multilingual automatic segmentation using German MAUS. In *Proc. of ICSLP 2000*, vol. 4, pp. 728–731, Beijing.

[2] Bloomfield, L. 1926. A set of postulates for the science of language. *Language*, 2: 153–164.

[3] Callan, D. E.; Tajima, K.; Callan, A. M.; Kubo, R.; Masaki, S.; Akahane-Yamada, R. 2003. Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *Neuroimage*, 19: 113–124.

[4] Chuu, C. 2003. LIESHOU: A Mandarin conversational task agent for the Galaxy-II architecture. Master's thesis, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, Cambridge; MA.

[5] Delmonte, R. 2000. SLIM prosodic automatic tools for self-learning instruction. *Speech Communication*, 30: 145–166.

[6] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In Strange, W., ed., *Speech perception and linguistic experience: Issues in cross-language research*, pp. 233–277. York Press, Timonium, Maryland.

[7] Pfitzinger, H. R. 1998. Local speech rate as a combination of syllable and phone rate. In *Proc. of ICSLP '98*, vol. 3, pp. 1087–1090, Sydney.

[8] 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.

[9] Pfitzinger, H. R. 2003. Acoustic correlates of the IPA vowel diagram. In *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 2, pp. 1441–1444, Barcelona.

[10] Pfitzinger, H. R. 2003. Using two- and three-dimensional fleshpoint measurements of articulatory kinematics in concatenative speech synthesis. In *6th Int. Seminar on Speech Production. Programme and Abstracts*, pp. 51, Manly, Australia.

[11] Tillmann, H. G.; Pfitzinger, H. R. 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. In *Proc. of ICSLP 2000*, vol. 3, pp. 295–297, Beijing.

[12] Tillmann, H. G.; Schiefer, L.; Pompino-Marschall, B. 1984. Categorical perception of speaker identity. In Broecke, M. P. R. v. d.; Cohen, A., eds., *Proc. of the Xth Int. Congress of Phonetic Sciences (Utrecht 1983)*, vol. IIB, pp. 443–448, Dordrecht.

[13] Trubetzkoy, N. S. 1939. *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague VII. Prag.

[14] Wu, Z. 2000. From traditional Chinese phonology to modern speech processing — realization of tone and intonation in standard Chinese. In *Proc. of ICSLP 2000*, vol. 1, pp. B1–B12, Beijing.

[15] Zierdt, A.; Hoole, P.; Tillmann, H. G. 1999. Development of a system for three-dimensional fleshpoint measurement of speech movements. In *Proc. of the XIVth Int. Congress of Phonetic Sciences*, vol. 1, pp. 73–75, San Francisco.