The Development of an Advanced SLP-based System for the Individual Learning and Fast Training of Speaking Skills in a New Foreign Language

Planning the Graduate College Research Program "Chinese for German Engineers (and other L1-speakers of German)"

Hans G. Tillmann & Hartmut R. Pfitzinger

Institut für Phonetik und Sprachliche Kommunikation
University of Munich, Schellingstrasse 3, 80799 München, Germany
tillmann@phonetik.uni-muenchen.de; hpt@phonetik.uni-muenchen.de

Abstract

This paper describes the organization of a concerted action of German and Chinese phoneticians, IT speech technologists and specialists in the field of foreign language teaching in order to create a new research program which has been proposed by the first author as a so-called *graduate college* (entitled "Chinesisch fr deutsche Ingenieure und andere L1-Sprecher des Deutschen") to the Federal Government of Germany.

The paper has four parts: We first start by describing the prime goals and conditions of our application-oriented basic research plans for the development of a new SL/T-system (Speech Learning/Training-system). Then we argue why — in phonetic speech research — the classic paradigm of "analysis-by-synthesis" should be replaced by a the programmatic approach which follows the new "synthesis-by-analysis"-principle. In the third part we look at certain types of teaching speech acts and teaching dialogs. In the last part we will be trying to consider the fact that such complex approaches of application-oriented basic speech research programs as in our case needs a concerted effort in the collaboration of all relevant disciplines within the field of SLP (including also NLP[1]), and here we end with the question of how the so-called dilemma of text-to-speech developers can be systematically avoided by practically involving the addressed users of the developed System (engineers and language students) from the very beginning.

1. Introduction

The authors are very grateful to the organizers of this conference for kindly having given us the opportunity to describe a research program that is still in its planning phase — so that it can also still be improved in some of its details before it becomes reality in the near future.

The application oriented aim of this new program is to organize a well coordinated basic research activity for producing the results that are needed for an advanced SLP-based automatic teaching system that really enables well-motivated individual L1-speakers of German to produce phonetically correct utterances of Standard Chinese after as short a training time as possible. Another important aim is to exploit — in this given application context — the IT-methods of modern speech technology for basic speech research. This aim is devoted to sharpening the instruments for investigating the actual functioning of natural speech acts — as far as the relation between speech acts and utterances is concerned. The final and prime aim of all our plans consists in contributing to the development of so-called *complete phonetic theories*[2] able to predict the phonetic form of any regular utterance of a given speaker in a given speech act.

It should be mentioned that some time ago the first discussion of a new research program was initiated after a delegation from the Institute of Acoustics and Language of the Chinese Academy of Science in Beijing had visited German speech research centers (in Munich IPSK, SIEMENS and in Berlin ZAS) and it was continued after we had got in closer contact to our phonetic colleagues of the Chinese

[1] As described by H. Fujisaki in his introduction to the First International Conference on Spoken Language Processing (ICSLP) in Kobe, 1990.

[2] The concept of a *Complete Phonetic Theory* (CPT) has been described in Tillmann & Pompino-Marschall 1993 [14] and Tillmann 1995 [11].

Academy of Social Sciences (CASS) in Beijing and the Fudan University in Shanghai. The initial partners in the German consortium of the project will consist of the Universities of Kiel, Berlin, Bonn, Saarbrücken, Stuttgart, München, the Technische Fachhochschule Berlin, the DFKI in Saarbücken and SIEMENS.

2. New Good Reasons for Applying the Munich Parametric High Definition (PHD) Speech Synthesis System

Inspired by the work of the pioneers in this new research field (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 [4] or the SLLS for Mandarin from MIT [3]) both authors were encouraged to proposed the application of the Munich PHD-System (described in Tillmann & Pfitzinger 2000 [12]) in a modified laptop-based version for the training of L1-speakers of German to teach them to reproduce — in a phonetically correct way — any given Mandarin Chinese learning items prompted by the system by just immediately presenting them two versions of their own utterance, namely (i) the original one actually produced by the learner when trying to categorically reproduce the teachers string of speech and (ii) a phonetically correct one, also in the speaker's own voice. This proposal had been based on the intuitively obvious, but yet unproved assumption that an immediate feedback of the directly comparable pair of the actually produced and (if necessary) phonetically corrected phonetic forms would improve the achievable results. This critical assumption has now been supported by the results of work from the Graduate School of Frontier Sciences at the University of Tokyo (Keikichi Hirose 2004 [8]).

For teaching non-Japanese learners to pronounce Japanese words with the correct accents Hirose first developed a method for automatically recognizing lexical accents (where each accent type was represented as a multidimensional Gaussian model). So the system could inform the learner whether the reproduction of the items presented in the teacher's voice were acceptable or not, in which latter case the learner's utterance was corrected in its prosodic form — using TD-PSOLA and taking the prosodic features from the teacher's prompting utterance as a model — and presented back to the learner together with a visual corrective feedback. In accent type training tests it could then be shown that this manner of giving corrective feedback improves the achievable results especially in the pronunciation of sentences (compared to giving feedback only in the voice of the teacher).

One important feature of the PHD-System implemented by the second author (Pfitzinger 2001 [9]) is that the speech signal of an utterance can be modified not only in its tonal F0-contour, but also in its local speaking rate contour. These PHD-techniques allow very flexible modifications of complex utterances. So it is not only possible to correct the tonal patterns of single syllables or words pronounced as citation forms, but also when these syllables are embedded and prosodically integrated into repeatable parts of fluent speech.

It should be said that these modifications can be realized in steps of increasing speech rates, and this can also be done with different degrees of tone-sandi as described in a famous invited paper (Wu 2000 [16]).

3. The proposed Synthesis-by-Analysis Approach as Opposed to the Classical Paradigm of Analysis-by-Synthesis

The idea of parametrically analyzing complex utterances and re-synthesizing the whole utterance with controlled modifications of selected parameter values has been relevant in Munich for the production of naturally sounding so-called real speech stimuli for perception experiments. So a proper parametric interpolation between phonemically identical utterances of two speakers delivered a continuum of absolute naturally sounding stimuli that could be used in identification and discrimination tests to demonstrate categorical perception of complex speaker identities (Tillmann et al. 1984 [15, 10]).

Both present authors, 2000 and 2004, have argued that analysis-by-synthesis speech research has been extremely helpful and productive for discovering elementary speech categories (including the trading relations which lead to a more complex picture), but that for the investigation of real speech the parametric modification of naturally produced complex utterances may lead to new and better insights concerning the variability of the phonetic facts that the speaking nervous system has to produce in any act of speech.

In the context of this paper we would like to add two further arguments in favour of the synthesis-by-

analysis research strategy in order to draw a conclusion that shows how important research on pronunciation teaching will be for the future development of phonetic speech science. Firstly complex categories such as words and prosodies in the voice of a given speaker — and not just only single elementary categories such as minimal pairs, phonemes, distinctive features, acoustic cues (i.e. VOTs etc.) are increasingly at the focus of phonetic speech science. The second argument says that any regular utterance, even the pronouncing of a minimal pair or a single vowel, is a very complex action, i.e. a complete act of speech. And in this context the great variability all of the complex, parametrically coded speech categories must be seen — as far as possible — in the light of systematic modification. With the idea of information bearing phonetic modification programmatically formulated in the title of a dissertation, experimental speech science has been born (Rousselot 1890).

4. The complex form of speech utterances in categorically different types of speech acts

In the context of phonetic modification the logically well established distinction between *autonymic* and *heteronymic forms of speech acts* receives a fundamentally new significance. There is no speech act without an utterance. And it is not only the text of an utterance but also the phonetic form of it that tells the hearer what kind of intention the speaker has in mind and wants to express.

We obviously find — in dialogs of pronunciation teaching a very typical class of autonymic speech acts where the speaker has the intention to demonstrate a categorically determined form, that is: the meaning of the utterance is the utterance itself in its categorically reproducible form. This can be the complex presentation of a single speech sound, the category of a tone on a certain vowel, or the citation form of a lexical item or a certain prosody to be put on a string of words.

Our research proposal implies that the SLP-technologies such as the PHD-System can be used to investigate which kinds of phonetic modifications express, or are related to, different intentions that are relevant in teaching the pronunciation of a new language. What modifications turn alphabetically explicit utterances of autonymically produced phonetic forms into those that we find as soon as the same words are used heteronymically in spontaneous speech.

Last but not least: which types and classes of modifications of all the different autonymically produced phonetic forms should be used in the teaching dialogs, and this again both in autonymic as well as in a heteronymic use and intention. There is an open list of unanswered questions. One of the most important problems to solve is to answer the question of phonetic correctness: in what teaching situations and intentions is what phonetic form produced by the learner acceptable or not.

5. The Research Strategy and Structure of a Distributed Graduate College

We believe that especially in our application-oriented basic research program it is good strategy to combine the interdisciplinarily oriented scientific education of beginning researchers with the possibility to start their own research with respect to the given list of questions that are a consequence of the application itself (Tillmann & Pfitzinger 2004 [13]). We cannot go into the details here, but it should be mentioned that graduates have to be trained in linguistics (including semantics and pragmatics), in phonetics, in DSP, in SLP but also to a certain degree in psychology and neurology.

The graduates will study in a distributed college at different institutes and universities in China and Germany. One prime question in all the PHD-work conducted by the students must be whether an SLP-based correction is acceptable and under what phonetic pronunciation conditions. The experience that we have made with the development of the MAUS-System (Beringer & Schiel 2000 [1]) to automatically segment and annotate with phonetic transcriptions spontaneous speech signals (given the orthographic text of the utterance) shows that verification of categories is possible where automatic recognition is yet impossible with today's ASR-technologies.

Quite another prime question would be to investigate whether and to what extent the von-der-Gabelentz-principle can be demonstrated, namely 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 and learn this language in a much shorter time (Georg von der Gabelentz 1891 [7]).

The third area of research will be devoted to the question how the SLP-based pronunciation training components can be integrated in larger NLP-based language learning dialogue systems such as those al-

ready mentioned.

The research strategy is determined by looking for teachable communicative relevant categories that can be successfully recognized, categorically modified and interpreted according to an intention expressed by a natural speaker in real speech. This also helps to avoid the well known dilemma of the developers of text-to-speech-systems who lose their ability to judge the speech quality of their own system.

6. 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] 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.

[3] 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; Massachusetts.

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

[5] Dioubina, O. I.; Pfitzinger, H. R. 2002. An IPA vowel diagram approach to analysing L1 effects on vowel production and perception. In *Proc. of ICSLP '02*, vol. 4, pp. 2265–2268, Denver.

[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] Gabelentz, G. v. d. 1891. *Die Sprachwissenschaft, ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig.

[8] Hirose, K. 2004. Accent type recognition of Japanese using perceived mora pitch values and its use for pronunciation training system. In *Proc. of the Int. Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages (TAL)*, pp. 77–80, Beijing.

[9] Pfitzinger, H. R. 2001. Phonetische Analyse der Sprechgeschwindigkeit. Forschungsberichte (FIPKM) 38, pp. 117–264, Institut für Phonetik und Sprachliche Kommunikation der Universität München.

[10] Pfitzinger, H. R. 2004. Unsupervised morphing between utterances of any speakers. In *Proc. of ICSLP '04*, pp. *submitted*, Korea.

[11] Tillmann, H. G. 1995. Three areas of Computational Phonetics. In *Proc. of the XIIIth Int. Congress of Phonetic Sciences*, vol. 4, pp. 72–75, Stockholm.

[12] 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.

[13] Tillmann, H. G.; Pfitzinger, H. R. 2004. Applying the Munich Parametric High Definition (PHD) speech synthesis system to the problem of teaching chinese tones to L1-speakers of German. In *Proc. of the Int. Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages (TAL)*, pp. 185–188, Beijing.

[14] 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.

[15] 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.

[16] 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.