Multi-Lingual Interoperability in Speech Technology,
RTO Meeting Proceedings 28, AC/323(IST)TP/4,
August 2000, Hull (Québec), Canada, pp. 15–19.

How foreign are "foreign" speech sounds? Implications for speech recognition and speech synthesis

Anders Lindström & Robert Eklund
{Anders.P.Lindstrom,Robert.H.Eklund}@telia.se
Telia Research AB, Farsta, Sweden

ABSTRACT

This paper reports results from a production study which shows in what ways the traditional Swedish phone set is expanded with phones similar to or approximating phones from other languages than Swedish in everyday speech. The inclusion of such sounds – here called *xenophones* – has implications for both automatic speech recognition and speech synthesis systems, especially in polylingual environments, which are discussed in the paper.

1. INTRODUCTION

In speech technology systems there is an increasing interest in issues such as dialectal variation, cross-language applications, handling of foreign accents et cetera. This problem is becoming more acute in an increasingly internationalized world, where people tend to speak more than one language, and also tend to ask for services that pay little or no attention to national or language borders.

A hitherto somewhat neglected problem that constitutes an important issue in the development of such multilingual applications is dealing with the fully normal inclusion of "foreign" speech sounds in the pronunciation of foreign names and words. Such speech sounds can be said to expand the phone inventory of the (native) language in question, a phenomenon observed in at least some languages, such as Swedish [5,6,7,10,11]. An example from Swedish would be the voiceless dental fricative [θ] (the first sound in the name "Thatcher"), which is not considered part of the Swedish phonemic inventory, but is nevertheless produced by approximately 50 percent of the population when pronouncing English words or names containing this sound in otherwise Swedish sentence contexts [6,10,11].

With a growing awareness of the need for multilingual automatic services (cf. e.g. [3]), the handling of language users' less constrained pronunciation becomes something of a *sine qua non*.

1.1. The Xenophone Problem

As was mentioned above, it has been shown that Swedes' pronunciation of names or words of foreign origin often exhibit sounds that are not part of what is considered the Swedish phoneme inventory. Such "added" sounds do not have a phonemic function in Swedish, and must therefore be attributed a particular status in the system. Even though they are not phonemes – or allophones of Swedish phonemes – they are clearly part of the *phone* sets of individual Swedish speakers. Hence, we suggested the term *xenophones* [6], i.e. "foreign phones", to denote such sounds.

Appropriate treatment of this phenomenon is likely to influence the performance of any speech recognition or synthesis system. For both these types of applications, expansions of the phone set are required. What is also apparent in the results reported in Eklund & Lindström [ibid.], is that the nature of this xenophonic expansion depends on the particular sound in question (among other things). This leads into the field of phonological acquisition, and more specifically, into the field of second language acquisition (SLA) research. The phonological processes involved when approaching a foreign language have been discussed in detail since long (e.g. [8,9]), and SLA research definitely provides valuable insight with regard to what factors might be at play. However, we would like to argue that although the phonological foundation is the same in xenophonic expansion and SLA, xenophones present a different problem since we are facing a different situation. Within SLA, the goal of the subject(s) is to master an entire target language, often in a target language context, whereas in the case of xenophonic expansion, the subjects simply include words of foreign origin in native-language sentences, mostly within fully native-language contexts. Thus, the entire communicative goal may be considered different, and this in turn should affect the actual rendering of the linguistic items in question. Indeed, as we shall see, this is supported by a some of our observations.

As discussed in Eklund & Lindström [7,10], a number of underlying factors can be assumed to be involved in governing the degree of adjustment. See Figure 1.

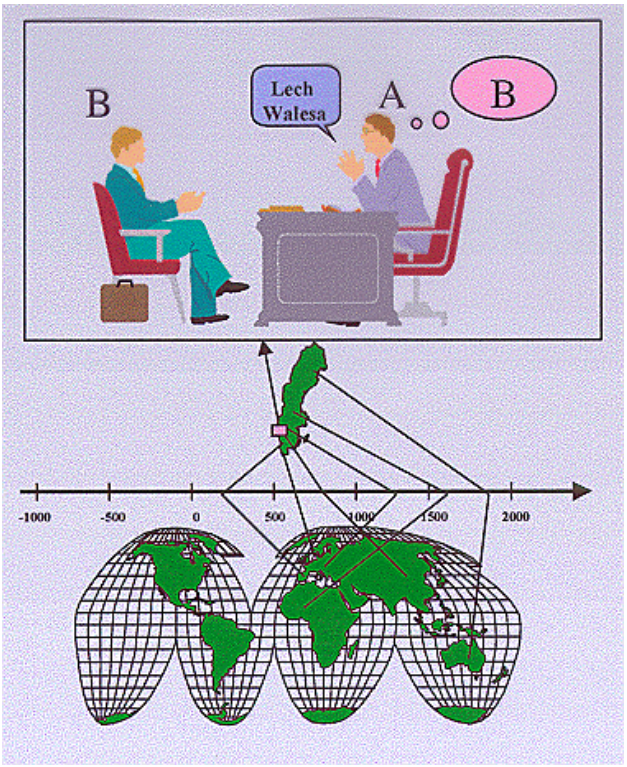


Figure 1: The language user in a typical situation. When speaker A is pronouncing a name in speaker B's presence, a number of factors are affecting the phonetic rendering of the name, such as the name's country of origin, the time it was introduced in speaker A's community, what channel it passed through, A's knowledge of B's language competence and other factors.

Multi-Lingual Interoperability in Speech Technology,
RTO Meeting Proceedings 28, AC/323(IST)TP/4,
August 2000, Hull (Québec), Canada, pp. 15–19.

These include – but are not limited to – the speaker's competence and performance capabilities with respect to the source language, the speaker's expectations of the listener's competence, the relative social status of speaker and listener, the socio-cultural distance to the country of origin, recency and frequency of the lexical item in question, and similarities and dissimilarities between the phonological systems involved.

2.2. Previous Work

Despite the fact that the problem is crucial, or even central in some languages, and despite the fact that references are found that date back to the 16th century, very little actual work on the phenomenon has been reported.

Maddieson [12] briefly discusses the phenomenon but simply refers to the phones in question (in several languages) as "anomalous" segments.

Abelin [1] discusses how to represent pronunciation of foreign (mainly English) words in *Svensk Ordbok*. She concludes that the English diphthongs [ɛɪ] and [ɔɪ] can be approximated with the Swedish sequences [ɛj] and [ɔj], respectively, but that the English diphthongs [əʊ] and [aʊ] are harder to accommodate. The English phone [z] is more or less always pronounced as [s] in Swedish, and the English alveolars [r, t, d, n] are normally realized as dentals in Swedish.

Eklund & Lindström [6] describe what English phones Swedes actually use in their speech, and show that a large proportion of Swedish speakers include "non-Swedish" sounds in their production system when pronouncing English words and names. Eklund & Lindström also describe the inclusion of xenophones into the Telia Research concatenative synthesizer.

Möbius et al. [13] mention that the German version of the Bell Labs multilingual TTS system has been augmented with phonetic units outside the German phone inventory in order to cover English and French speech sounds.

2. METHOD

In order to acquire information and knowledge concerning Swedish speakers' usage of xenophones, and also, to some extent, insight in their expectations on xenophone usage, a production study was conducted. The rationale for looking at production data, we argue, is that knowledge may be gained in several dimensions: Which English phones have an effect of the Swedish subjects' productions? What is the nature of this effect—is the phone repertoire extended or does some kind of segmental mapping take place? Even if a speaker does not produce an English name or word in an accent-free manner, he or she might still do something that clearly lies outside the Swedish phone inventory. By producing something that is neither Swedish nor English, as it were, the speaker is indicating an awareness of the difference between the English pronunciation and a fully rephonematized pronunciation (i.e., "translating" the English sound into its phonetic "counterpart" in Swedish). This provides important information in the "attitude dimension", insofar as it shows that even speakers who do not fully master the production of English sounds might expect these sounds to occur in particular words.

2.1. The Linguistic Material

A set of twelve sentences was constructed containing the 15 English speech sounds [tʃ, dʒ, ʃ, ʒ, θ, ð, z, , ɬ, w, aɪ, eɪ, əʊ, juː, æ]. The two non-English (and non-Swedish) sounds [x,aː] were also included in the material. All these sounds were chosen so that they would differ phonetically from Swedish speech sounds to varying degrees, and so that none of them would be included in any traditional description of the Swedish phonological system.

The phones were included in commonly known names and words in twelve fully natural Swedish sentences and it was assumed that the words and names in which the xenophones appeared would be known by the bulk of the subjects.

Two example sentences from the material are given below.

Många har Roger Moore som favorit i rollen som James Bond.
("A lot of people prefer Roger Moore's interpretation of James Bond")

Intercity-tåget gick direkt från Aachen till Baden-Baden.
("The Intercity train went straight from Aachen to Baden-Baden")

2.2. Recordings and Subjects

The sentences were included in a much larger session of linguistic material recorded to train the Telia/SRI Swedish speech recognizer as a part of the *Spoken Language Translator* (SLT) project [2,14]. The material was presented under the heading 'Kändisar' (Celebrities), and it can be assumed that subjects were unaware of the fact that their pronunciation was the object of study.

The subjects were all Telia employees or relatives of Telia employees. The age span was 15 to 75. Hi-fi recordings were obtained of more than 460 subjects on 40 different locations covering the whole of Sweden, so that data from all major dialect areas were obtained. In this way a total of approximately 29,000 xenophone tokens were collected. The subjects also filled in forms, providing information concerning educational level, regional origin and so on.

2.3. Evaluation

Three phonetically trained native speakers of Swedish, with an above-average knowledge of English, transcribed the target phones, using a fairly narrow allophonic transcription scheme. So far, 15,202 potential xenophone tokens have been evaluated.

3. RESULTS

Figure 2 shows the proportional distribution of the subjects' productions of the speech sounds [aː, aɪ, eɪ, əʊ, juː, æ, tʃ, dʒ, x, ʃ, ʒ, θ, ð, z, ɚ, ɬ, w], where each instance of these has been assigned to one of three categories along the awareness and fidelity dimensions.

Category 1 corresponds to high awareness coupled with high fidelity, production-wise.

Category 3 indicates low fidelity, and probably low awareness, although it may also be the case that some speakers deliberately rephonematize (for normative reasons).

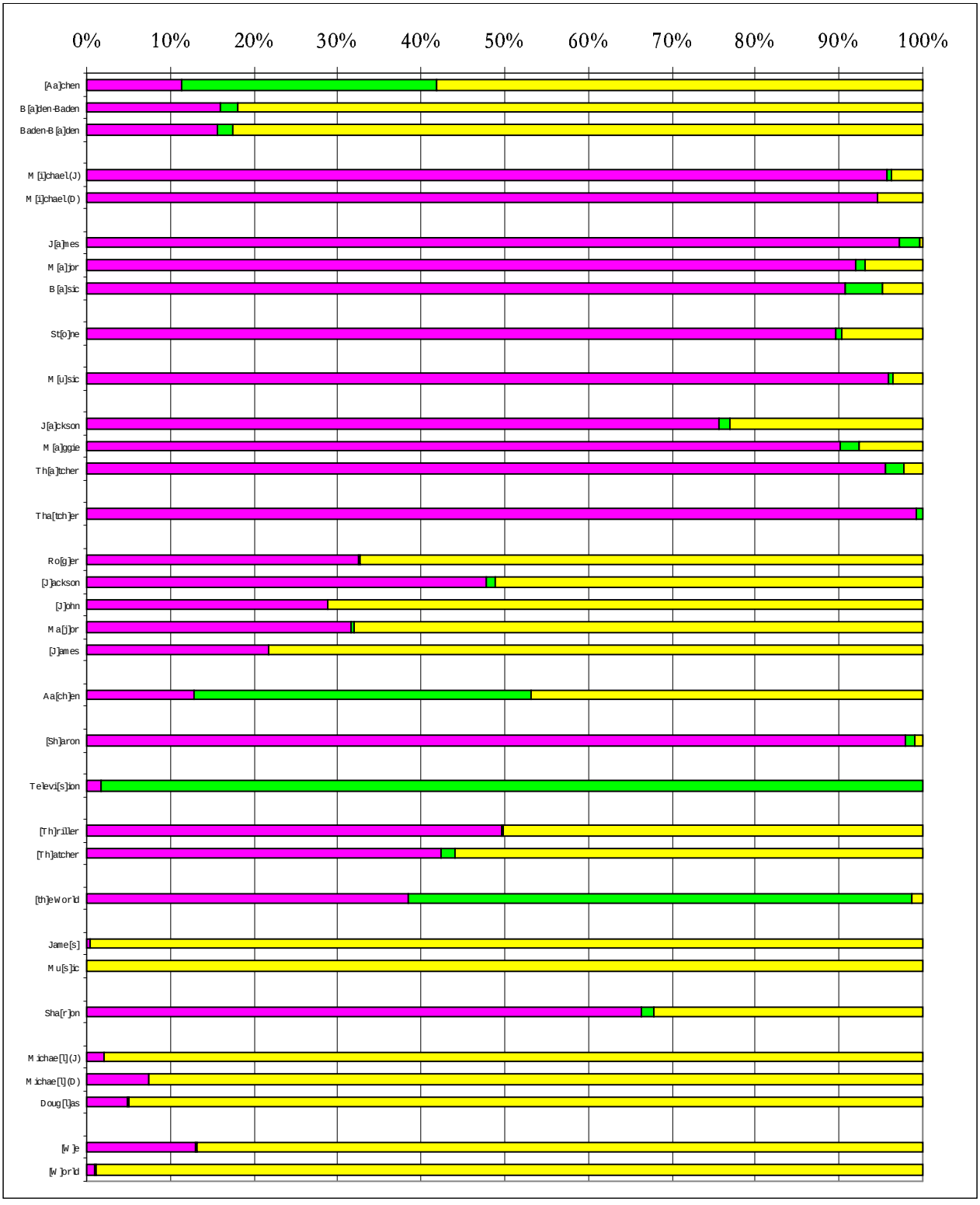
Category 2, high awareness and low fidelity, is interesting, since it represents those speakers who are apparently aware that *something* foreign should be going on, but fail to produce a good enough approximation of the "target" speech sound. Speakers in this category can certainly cause considerable problems for ASR systems.

As can be seen in Figure 2, the distribution over the three categories differs considerably as a function of target phone, and even as a function of each individual "lexical item". It is interesting to note that voiced fricatives are more or less non-existing, despite the fact that are easy to produce, whereas the more "remote" phones (from a number-of-phonetic-features perspective), from a Swedish point of view, e.g. dental fricatives, are produced by a large number of subjects.

A subset of the data presented here has also been evaluated with respect to which underlying factors might explain the differences in use of xenophones.

Multi-Lingual Interoperability in Speech Technology,
RTO Meeting Proceedings 28, AC/323(IST)TP/4,
August 2000, Hull (Québec), Canada, pp. 15–19.

Figure 2: For each target English speech sound and each occurrence in the read sentences, the proportional distribution of the Swedish subjects' productions is shown. Based on the similarity between the produced sound and the target phone, the different productions are assigned to one of three categories along two dimensions, the *awareness dimension* (to what extent people are aware of the difference between Swedish and English pronunciation), and the *fidelity dimension* (how well they succeed in the production of the foreign sounds). The first category (magenta/dark grey) corresponds to a high awareness among the subjects coupled with a high capability in rendering a sound close to the one in the source language. The second category (green/middle grey) corresponds to the case where the subjects were apparently aware that something "non-Swedish" would be appropriate, but failed to produce a good approximation. The third category (yellow/light grey) corresponds to full adjustment to Swedish.



Lindström and Eklund [11] showed that age seems to be one factor that systematically affects the productions, in such a way that the youngest and oldest subjects generally produce relatively more category 2 and 3 productions than do the other subjects.

In the same study, no significant gender differences were found, nor were there any systematic regional differences. The last result, however, may be due to lack of control for the variable "educational level", and re-evaluation of the data with that in mind is in the works.

4. DISCUSSION

Xenophones can be discussed and studied from several different angles. From a theoretical perspective, the underlying theoretical issues xenophones raise mainly concern general phonological acquisition, relating to, without being similar to, SLA research.

As indicated in Figure 1, there are a number of underlying factors that can be assumed to be at play in determining the choice of the speaker's pronunciation strategy, and we believe that we have shed some light on the issue of what speakers do when solving this task of finding the socially acceptable level along the awareness/fidelity dimension.

From a theoretical side, the "foreignness" of such sounds can be discussed. If most Swedes use certain sounds in everyday conversation, and/or expect them to be used, how "foreign" are they in the language community? Moreover, in a world that is characterized by increasing international communication – economical, cultural, social – such cross-breeding between languages can be expected to become more and more frequent.

From a more practical side, there are a number of consequences that these observations are bound to have for automatic speech recognition and speech synthesis.

4.1. Implications For Recognition

A recognizer is facing the entire variety of speech sounds within a given speech community, and the modelling of what it can be expected to hear boils down to a few crucial issues.

First, the standard view on what the Swedish phone set looks like must be reconsidered, since it obviously to a large degree contains sounds normally not considered "Swedish", despite the fact that a large number of Swedish speakers do use them in normal conversation.

To complicate matters further, a word/name of foreign origin and containing foreign – or foreign-similar – sounds can appear in an otherwise Swedish sentence, which means that the recognizer needs to handle phones from (at least) two languages at once. Within the SLT project, a recognizer that is able to handle English and Swedish was developed [4,15,16]. The recognizer is capable of recognizing the odd Swedish word inside an otherwise English sentence, and vice versa.

Another issue is exactly how acute a problem xenophones present to a recognizer. This, of course, depends heavily on the context and discourse. An application like automatic handling of film ticket purchasing would surely need to cope with a large number of xenophones, since most English film titles are not translated into Swedish. Within other domains, such as bookings of summer houses in the Stockholm archipelago, xenophones are not likely to occur at all. Thus, xenophone inclusion for a given application is also an empirical issue.

4.2. Implications For Synthesis

As opposed to recognition, where the entire variety needs be considered and catered for, a synthesizer probably only needs to cover *one* acceptable variety. The operative word here, of course, is "acceptable". Although it is our belief that a production study provides information in the acceptability domain insofar as it can be assumed that users of speech synthesis systems will be less prone to accept a synthesizer with a lower level of competence than themselves, the only safe method to gain insight in the acceptability domain would be to conduct a perception study. One such method would be to play back to subjects the obtained recordings and ask them rank the pronunciations along a few dimensions, such as intelligibility, "intelligence", pleasantness and so on. It is our belief that a low inclusion level of xenophones might not primarily show up in the intelligibility dimension, but rather present itself to listeners as a synthesizer with a low educational level.

Another problem to consider is that "maximizing" in the xenophone dimension might leave certain listeners behind, especially concerning languages that are not so commonly known as English (e.g. French, German or Russian) and that an appropriate level must be found. It can be assumed that choosing too "high" a level will signal an attitude which would be perceived as high-browed and obnoxious. This, too, needs more studies.

To the best of our knowledge, few attempts to include xenophones in synthesizers have so far been made. As mentioned above, Eklund & Lindström [6] report the inclusion of English xenophones in the Telia Research research synthesizer and Möbius et al. [13] mention the inclusion of a few English and French sounds in the German version of the Bell Labs multilingual TTS system.

4.3. Future Research

Apart from the perception studies mentioned above, a deeper look into the phonological-regional dimension is needed. The rationale for doing this is that one thing one would want from an intelligent recognizer is that it possess a certain level of predictive power, so that it could "tune in" to a particular speaker's use of xenophones (and idiosyncratic speech behavior in general). However, our observations so far do not provide much hope in that dimension, since the speakers generally do not exhibit a high degree of consistency in their use of xenophones. For example, a phrase like *Diana and Charles* (from the material) may be pronounced with xenophones on *Diana* but not on *Charles*, or vice versa. Thus, our studies so far indicate that xenophone inclusion may appear spot-wise, rather than consistently. However, this asks for more research.

Another thing that awaits studies is to what extent prosodic signaling is employed. Some subjects signaled awareness of the foreignness of the names and words by using a prosodic realization that is influenced by the source-language, in this case English, either in addition to, or independently of, the use of xenophones. So far, we have not conducted any formal studies of this phenomenon, and the benefits from such knowledge of course require that recognizers make use of prosody, something which currently is not done, at least not to any larger degree.

Another factor to be studied further is the role of orthography, something we have tried to normalize for by including the same sounds with different spelling (i.e., the voiced affricate [dʒ] was presented both in the name *James* and in the name *Roger*). It proved to have some effect [6], but more data are needed before any far-reaching conclusions made be drawn concerning the role of orthography.

Finally, an obvious factor to study is the speakers' educational level. It goes without saying that previous and close familiarity of foreign languages affect the pronunciation, as well as one's expectations on how names and words of foreign origin "should" be pronounced. Such studies are underway, and will be reported in future work.

5. ACKNOWLEDGEMENTS

The authors thank Per Sautermeister for help with the listening task. Also thanks to Malin Ericson, who made comments on draft versions of the paper.

6. REFERENCES

1. Abelin, Å. 1985. Om uttalsmarkering och uttalsregler i svensk ordbok. *Rapporter från Språkdata (21)*, Gothenburg University, Department of Computational Linguistics, Gothenburg.
2. Becket, R., P. Boullion, H. Bratt, I. Bretan, D. Carter, V. Digalakis, R. Eklund, H. Franco, J. Kaja, M. Keegan, I. Lewin, B. Lyberg, D. Milward, L. Neumeyer, P. Price, M. Rayner, P. Sautermeister, F. Weng & M. Wirén. 1997. *Spoken Language Translator: Phase Two Report*. Telia Research AB and SRI International.
3. Billi, R. n.d. Interview published in *Le Journal – The Journal of Record for Human Language Technology*. http://www.linglink.lu/lejournal/article.asp?articleIndex=628
4. Digalakis V. & L. Neumeyer. In press. Multiple Dialects and Languages. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator*. Ch. 18, pp. 307–318. Cambridge University Press.
5. Eklund, R., J. Kaja, L. Neumeyer, F. Weng & V. Digalakis. In press. Porting a Recognizer to a New Language. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator*. Ch. 17, pp. 297–306. Cambridge University Press.
6. Eklund, R. & A. Lindström. 1998. How To Handle "Foreign" Sounds in Swedish Text-to-Speech Conversion: Approaching the 'Xenophone' Problem. *Proc. of ICSLP 98*, Sydney, November 30–December 5. Paper 514, Vol. 7, pp. 2831–2834. CD-ROM available from Causal Productions Pty Ltd, PO Box 100, info@causal.on.net.
7. Eklund, R. & A. Lindström. 1996. Pronunciation in an internationalized society: A multi-dimensional problem considered. *FONETIK 96, Swedish Phonetics Conference, Nässlingen, 29–31 May, 1996. TMH-QPSR 2/1996*, 123–126.
8. Flege, J.E. 1987. Effects of Equivalence Classification on the Production of Foreign Language Speech Sounds. In James, A. & J. Leather (eds.). *Sound Patterns in Second Language Acquisition*, Foris Publications.
9. Hammarberg, B. 1990. Conditions on Transfer in Second Language Phonology Acquisition. In Leather, J. & A. James (eds.), *New Sounds 90, Proc. of the 1990 Amsterdam Symposium on the Acquisition of Second-Language Speech*. University of Amsterdam.
10. Lindström A. & R. Eklund. 1999. Xenophones Revisited: Linguistic and other underlying factors affecting the pronunciation of foreign items in Swedish. *Proc. of ICPhS 99*, San Francisco, August 1–7. Paper 0708.
11. Lindström A. & R. Eklund. 1999 [jɑːmes] or [dʒɛɪmz] or Perhaps Something In-between? Recapping Three Years of Xenophone Studies. Gothenburg Papers in Theoretical Linguistics, 81. *Proc. Fonetik 99*, The Swedish Phonetics Conference, June 2–4 1999, pp. 109–112.
12. Maddieson, I. 1984. *Patterns of sounds*, Cambridge University Press.
13. Möbius, B., R. Sproat, J.P.H. van Santen & J.P. Olive. 1997. The Bell Labs German Text-to-Speech System: An Overview. *In Proc. ESCA. Eurospeech 97, Rhodes, Greece*, ISSN 1018–4071, pp. 2443–2446.
14. Rayner, M., D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). In press. *The Spoken Language Translator*. Cambridge University Press.
15. Weng, F. In press. Language Modeling for Multilingual Speech Translation. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator*. Ch. 16, pp. 281–296. Cambridge University Press.
16. Weng, F., H. Bratt, L. Neumeyer & A. Stolcke. 1997. A Study of Multilingual Speech Recognition. *Proc. Eurospeech*, pp. 359-362, Vol. 1, Rhodes, Greece.