How To Handle "Foreign" Sounds in Swedish Text-to-Speech Conversion: Approaching the 'Xenophone' Problem

Robert Eklund & Anders Lindström

Telia Research AB, Farsta, Sweden

ABSTRACT

This paper discusses the problem of handling "foreign" speech sounds in Swedish speech technology systems, in particular speech synthesis. A production study is made, where it is shown that Swedish speakers add foreign speech sounds, here termed 'xenophones', to their phone repertoire when reading Swedish sentences with embedded English names and words. As a result of the observations, the phone set of a Swedish concatenative synthesizer is extended, and it is shown (by example) that this produces more natural-sounding synthetic speech.

1. INTRODUCTION

In recent years, both automatic speech recognition (ASR) and text-to-speech conversion (TTS) systems have attained quality levels that allow inclusion in every-day applications. This does not mean, however, that all problems with regard to ASR/TTS are solved. The particular problem addressed in this paper is the fact that in a language such as Swedish, what is normally regarded as the Swedish phone inventory is quite often expanded with phones from other languages, notably English. As a consequence, in order to create high-quality ASR and TTS systems for Swedish, these "foreign" phones need to be included in the underlying language description. However, although this problem per se is uncontroversial in that several researchers acknowledge the existence of "non-Swedish" sounds in every-day spoken Swedish, to the best of our knowledge, there have been no formal studies to establish exactly what this extended phone set looks like. In this paper, we will first attempt to determine the nature, including phone frequencies, of this extended phone set when dealing with words and names of English origin, and then show how this can be directly applied to enhance the quality of our Swedish TTS system, by extending its phone set.

1.1. The 'Xenophone' Problem

A word or name of foreign origin can be pronounced with varying degrees of adjustment to the Swedish phonological system. This variation spans the whole range from virtually no adjustment (i.e. the pronunciation is close to that of the source language), via some degree of rephonematization, to total adjustment (for instance orthographically oriented, 'naïve', pronunciation). As discussed in Eklund & Lindström [3], a number of underlying factors can be assumed to be involved in governing the degree of adjustment, including (but 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/dissimilarities between the two phonological systems in question. It should be noted that since the dialects of a language differ in terms of phonology, it would not be surprising to find dialect-specific variation in the treatment of foreign items, because of this last factor.

Within the somewhat related field of second language acquisition, SLA, a central problem is how speakers of a native language (L1) approach a foreign, target language or second language (L2). It is argued that when a language learner perceives a sound in L2 as sufficiently close to a sound in the L1, they pass for the same sound, and are, in effect, put in the same 'equivalence class'. The effect of this equivalence classification has been shown by e.g. Flege [4]. Hammarberg [5], points out that whether or not an L2-sound is perceived to be identical or similar to an L1-sound is not a yes/no-decision, but is a gradual phenomenon that depends on several factors, such as the naturalness of the L1 sounds per se (in markedness terms), and/or the learner's current level of competence in L2. A consequence of this equivalence classification is that L2 sounds that lack similar items in L1 will be learned faster, since they are more easily perceived as different. Although similar in some respects, SLA research does not deal with exactly the same problem we are discussing here. The focus of SLA research is to study how speakers of a language approach an L2 with the intention to master a substantial part of it, sometimes while also living in an L2 community. In the case discussed in this paper, lexical items from another language appear in utterances in the native language, in this case Swedish, and what we are interested in is to see what type and degree of adjustment to the native language is normally applied.

Sounds that are foreign to the phone inventory of a language have a special status in the phonological system of that language. On the one hand, one cannot claim that they be a part proper of that system, but on the other hand, they might have such a status that most people would expect them to be used in certain linguistic contexts. Maddieson [6] calls these sounds 'anomalous'. Another term encountered for these sounds is 'loan phonemes'. Since the former term is not very clear with regard to meaning, and the latter is somewhat dubious since it is far from clear that we are dealing with phonemes proper, we will in the following use the term 'xenophones' (meaning 'foreign sounds') to denote such sounds.

1.2. Previous Work

The problem of accommodating foreign words or names in general is far from new, and several Swedish and German references on the subject date back to the 16th century. In more recent literature, Abelin [1] discusses how to represent pronunciation of foreign (mainly English) words in *Svensk Ordbok*. She concludes that the English diphthongs [eɪ] and [oɪ] can be approximated with the Swedish sequences [ej] and [oj], 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.

Möbius et al. [7] 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.

In 1996, in an earlier attempt to address the problem discussed in this paper, Eklund & Lindström [3] investigated what English phones Swedes actually use in their speech. It was shown that a large proportion of the speakers included "non-Swedish" sounds in their production system, and used them when pronouncing English words and names. As a consequence of the 1996 study, a set of polyphones, modified to encompass some of the xenophones studied, was recorded later that year for inclusion in a concatenative synthesizer. A large pronunciation dictionary including proper names of foreign origin was also produced. The resulting speech synthesis, however, was not evaluated in the paper. Moreover, the 1996 study was based on 70 speakers only (35 from Stockholm and 35 from Scania). In order to eliminate the risk of regional bias in the material it was decided to study the phenomenon in most major Swedish dialects. This could also lay the foundation for future studies of the differences between dialects with regard to their underlying phonological structure.

2. METHOD

This study is a continuation of the 1996 study previously mentioned. By looking at production data, insight 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 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 Swedish pronunciation. 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, ɚ, ɬ, w, aɪ, eɪ, əʊ, juː, æ], which were chosen because they were judged to be possible candidates for the processes described in the previous section. The chosen sounds differ phonetically from Swedish speech sounds to varying degrees. The Swedish phonological system is normally not described to include any of the above sounds, but the following remarks could be made: The Swedish retroflex [ʂ] (which is lacking in Southern varieties of Swedish) is phonetically quite close to an English [ʃ], but there is no voiced postalveolar or alveolar fricative in Swedish, and neither a voiced nor an unvoiced dental fricative. There is a voiced retroflex fricative in Swedish ([ʐ]), which is an allophone of /r/. The voiceless affricate could possibly be approximated by Swedish [t] + the alveolo-palatal fricative [ɕ], pronounced in sequence, but there is no voiced counterpart. Swedish /l/ is normally a lateral approximant, but in Northern varieties of Swedish, often velarized. The approximant [w] lacks correspondence in Swedish, although a similar sound may appear as a final element in diphthongized rounded vowels. Of the vowels and diphthongs chosen, [aɪ, eɪ, juː] could quite easily be approximated, using Swedish [j] combined with [a, e, uː], as Abelin suggested. Her suggestion concerning [aʊ] was disregarded, since that diphthong must be considered internalized in Swedish in words such as *aula* and the prefix *auto-*. Finally, [æ] appears as an allophone in Swedish preceding /r/ and retroflex consonants.

The twelve sentences included names and words of English origin that were deemed to be commonly known, embedded in a Swedish sentence in a natural way. An example is shown below:

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

2.2. Data Collection and Evaluation

The sentences were included in a much larger session of linguistic material recorded to train the Telia/SRI Swedish recognizer [2], which was also the main purpose of the data collection. The sentences were presented under the heading 'Kändisar' (Celebrities). Thus, it can be assumed that subjects were unaware of the purpose of the recordings, i.e. they did not know 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. The sentences were recorded by more than 460 subjects in 40 different locations covering the whole of Sweden. Thus, all major dialects were covered. In this way a total of 13,343 tokens were collected.

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. It was a deliberate decision not to use native speakers of any English variety as transcribers, since we were not so much interested in which productions sound English to an Englishman, as in what sounds non-Swedish, or too Swedish, to Swedish people. The transcribers also made note of sentences where the subjects applied total adjustment, using exclusively Swedish allophones in their pronunciation of the foreign items.

3. RESULTS

Preliminary results, drawn from the production study, show that very few of the subjects (less than a dozen) resorted to total re-phonematization. Instead, the majority of the subjects expanded their allophonic repertoire considerably, despite the fact that the foreign items were embedded in a Swedish context in a fully plausible way. The results are shown in Table 1a (vowels) and Table 1b (consonants).

All the target vowels (except [æ] in the name *Jackson*) and diphthongs are very well approximated in 90 % of the cases or more, remarkably enough also for the diphthong [əʊ], which is quite dissimilar from any 'normal' Swedish vowel. The results for the consonants indicate that the subjects almost without exception produced the voiceless affricate [tʃ], while the figures for the voiced counterpart [dʒ] ranged from 21 % in *James* to 48 % in *Jackson*, which is also quite remarkable, since there is normally no such thing as a voiced affricate in the Swedish phonological system. The retroflex fricative [ʂ] that 60 % of the subjects produced in *Sharon* could be regarded as a sufficient approximation of the postalveolar [ʃ], as could perhaps also the alveolo-palatal fricative [ɕ], produced by 32 % of the subjects. More detailed analysis show that of the Scanian subjects, 90 % produced the alveolo-palatal fricative, which is not surprising, since Southern Swedish lacks the retroflex. Both the voiced [ð] and the unvoiced [θ] dental fricative was produced to an amazingly high degree, considering the lack of similar speech sounds in Swedish, while virtually no subjects succeeded in producing the voiced alveolar [z] and postalveolar [ʒ] fricative. Subjects also chose to opt for almost full adjustment to Swedish in the case of [l], and to some extent also in the case of [w].

Another observation that was made was that subjects were not at all consistent across lexemes. In pronouncing “Roger Moore” and “James Bond”, the same subject would produce an affricate on the first target phone, but not on the second, or vice versa. This might make it difficult to create a hierarchy of sound accommodation.

Table 1a: For each target English speech sound and each occurrence in the read sentences, the resulting distribution of the Swedish subjects’ productions, as obtained by manual phonetic transcription, is shown as a percentage. 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 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 corresponds to the case where the subjects were apparently aware that something “non-Swedish” would be appropriate, but failed to produce a good approximation. Finally, the third corresponds to full adjustment to Swedish. Cases where the transcribers were unable to hear what was produced are marked by an asterisk.

		Category	***1***		***2***		***3***	
		Awareness	***high***				***low***	
		Fidelity	***high***		***low***			
Target	***No. of tokens***	***Occurring in the word***						
aɪ	456	**M**ichael	95.8	aj	0.4	ej	3.1	iː
		(Jackson)					0.7	ɪ
	460	**M**ichael	94.6	aj			3.0	iː
		(Douglas)					2.4	ɪ
eɪ	465	**J**ames	97.2	ej	2.4	ɛ	0.4	a
	463	**M**ajor	92.0	ej	0.9	ɛ	6.9	a
					0.2	aj		
	460	**B**asic	90.7	ej	3.5	ɛ	3.9	a
					0.4	ɛj	0.9	aː
					0.4	æː		
					0.2	ɔ		
əʊ	460	**St**one	89.2	əʊ	0.4	oːʊ	5.0	oː
					0.2	ɔuː	4.4	uː
					0.2	ɪuə	0.2	ɔ
					0.2	uːə		
					0.2	*		
juː	452	**Mu**sic	96.0	juː	0.2	jʉː	3.6	ʉː
					0.2	jəj		
æ	456	**J**ackson	75.7	æ	1.1	ɛ	23.0	a
					0.2	*		
	463	**M**aggie	90.2	æ	0.6	ɛ	7.1	a
					1.5	*	0.6	aː
	463	Th**a**tcher	95.5	æ	1.7	i(ː)	2.2	a
					0.2	ej		
					0.2	ɛ		
					0.2	*		
Total	4598							

Table 1b: Consonants. (See Table 1a for an explanation.)

		Category	***1***		***2***		***3***	
		Awareness	***high***				***low***	
		Fidelity	***high***		***low***			
Target	***No. of tokens***	***Occurring in the word***						
tʃ	463	Tha**tch**er	99.2	tɕ	0.4	ʂ		
					0.2	ɕ		
					0.2	*		
dʒ	465	Ro**g**er	32.5	dʒ	0.2	gd	67.3	g
	456	**J**ackson	47.8	dʒ	0.4	di	51.2	j
					0.2	ɕ		
					0.4	*		
	463	**J**ohn	28.9	dʒ			71.1	j
	463	Ma**j**or	31.6	dʒ	0.4	*	68.0	j
	465	**J**ames	21.7	dʒ			78.3	j
ʃ	460	**Sh**aron	60.1	ʂ	0.4	ʃj	0.9	s
			31.5	ɕ	0.4	tɕ		
			6.3	ʃ	0.2	sɕ		
					0.2	*		
ʒ	452	Televi**s**ion	1.7	ʒ	94.6	ʂ		
					3.3	ʃj		
					0.2	ɕ		
					0.2	ʂʃj		
θ	456	**Th**riller	49.6	θ	0.2	ɕ	48.0	t
							2.2	t̪
	463	**Th**atcher	42.3	θ	1.3	tɕ	56.0	t
					0.2	s		
					0.2	*		
ð	462	**th**e World	38.5	ð	57.5	d	1.3	t
					1.1	r		
					0.4	h		
					0.4	v		
					0.2	j		
					0.2	0		
					0.4	*		
z	465	Jame**s**	0.4	z			99.6	s
	452	Mu**s**ic					100.0	s
ɚ	460	Sha**r**on	62.0	z̩	0.4	ʁ	32.2	r
			4.4	ɚ	0.2	d		
					0.2	n		
					0.2	t		
					0.4	*		
ɬ	456	Michae**l** (J)	2.0	ɬ			98.0	l
	460	Michae**l** (D)	7.4	ɬ			92.6	l
	460	Doug**l**as	4.8	ɬ	0.2	0	95.0	l
w	462	**W**e	13.0	w	0.2	m	86.8	v
	462	**W**orld	0.9	w	0.2	b	98.9	v
Total	8745							

There may well also be graphemic influence on the data. The affricate [dʒ], when spelled with a <g> might behave quite differently from [dʒ], spelled with a <j>, but our limited material does not allow us to draw any further conclusions.

A final observation is that a few speakers quite obviously changed 'mode of speaking', that is to say, they could start out with a Swedish pronunciation of "Roger", but realizing when arriving at "Moore" that it was not the Swedish name, in which case they backed to restart with a more English pronunciation.

4. TTS IMPLEMENTATION

As was mentioned previously, a set of xenophone polyphones was recorded in 1996 based on the aforementioned study. The recorded items have now been tested in synthesized speech. The synthesizer in question is a concatenative synthesizer, using a female voice, developed at Telia Research AB, in the Spoken Language Processing laboratory. The basic unit is the demi-syllable, with the addition of some other items, such as derivational endings (not already covered), closed word classes, nasal triphones and some other items. The current set of such polyphones counts around 15,000 units, including the poly-phones containing xenophones.

When encountering an English word or name, a Swedish TTS system is facing the options of either trying to do the best it can with a purely Swedish set of polyphones, or make use of the added xenophone polyphones. From our study, it can be concluded that most Swedish speakers do seem to expect something outside a Swedish rendering of such lexical items. To illustrate this, a set of sentences (the same set as was used in the data collection) was synthesized in two ways: First using Swedish polyphones only (trying to make the best of it) and second, making use of the added xenophone polyphones. Two examples are shown below (xenophones marked in boldface):

(a) *Det anses allmänt att **J**ohn M**aj**or är en blek efterträdare till M**a**ggie **Th**a**tch**er* [0514_01.WAV]
("It is a widely held opinion that John Major is a rather pale successor to Maggie Thatcher")

(b) *Många rockstjärnor medverkade i sången "**W**e are **th**e **W**orld"* [0514_02.WAV]
("Many rock stars participated in the song 'We are the World'")

Preliminary, informal evaluation shows that using the added xenophone polyphones produces speech output much more natural-sounding, in accordance with the data in our study.

5. DISCUSSION AND FUTURE WORK

Whereas some of the xenophones, as mentioned above, could easily be approximated by using Swedish phones, the results of our study indicate, in a quite convincing way, that the phone set of any Swedish synthesizer would need to be extended to encompass at least [dʒ, θ, ð, əʊ] and possibly [w], since that supposedly represents a "lower limit" with regard to what Swedish listeners would expect from a Swedish speech synthesizer. A Swedish recognizer, on the other hand, would have to incorporate an even larger set to cover the production variability observed in this study.

In approaching the problem of xenophones there are several factors that need to be considered. The approach opted for in this study is based on production rather than perception or evaluation. The rationale behind choosing a production-based approach is that, we argue, it shows people's attitudes towards the occurrence of foreign items in a more subconscious way than if they were told to evaluate the quality of different versions of synthesized speech. The method applied in this paper provides information in at least 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). In this way, we have managed to get an impression of both to what degree Swedish listeners expect these items to be given non-Swedish realizations, as well as some hints with regard to how well Swedish listeners want the said items to be pronounced. It must be mentioned, however, that one problem associated with this method is that we cannot know whether we are testing language, word or world knowledge.

What is not studied at all, or only to a very limited degree so far, is what rôle social background, age, gender and regional background might play here. Whereas social background lies beyond what can be deduced from the material, a closer study would show what differences there are concerning the other factors mentioned above. It would also be interesting to include word prosody in the production study. Another object of further study is to what degree the same assumptions are valid for borrowed items from other languages.

6. ACKNOWLEDGEMENTS

The authors thank Per Sautermeister for help with the listening task. We are also grateful to Catriona MacDermid and Mats Wirén for comments on draft versions of this article.

7. REFERENCES

1. Abelin, Å. Om uttalsmarkering och uttalsregler i svensk ordbok. *Rapporter från Språkdata (21)*, Gothenburg University, Dept. of Computational Linguistics, Gothenburg, 1985.
2. Becket, R., Boullion, P., Bratt, H., Bretan, I., Carter, D., Digalakis, V., Eklund, R., Franco, H., Kaja, J., Keegan, M., Lewin, I., Lyberg, B., Milward, D., Neumeyer, L., Price, P., Rayner, M., Sautermeister, P., Weng, F. & Wirén, M. *Spoken Language Translator: Phase Two Report*. Telia Research AB and SRI International, 1997.
3. Eklund, R. & Lindström, A. 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, 1996.
4. Flege, J.E. Effects of Equivalence Classification on the Production of Foreign Language Speech Sounds. In James, A. & Leather, J. (eds.). *Sound Patterns in Second Language Acquisition*, Foris Publications, 1987.
5. Hammarberg, B. Conditions on Transfer in Second Language Phonology Acquisition. In Leather, J. & James, A. (eds.), *New Sounds 90, Proc. of the 1990 Amsterdam Symposium on the Acquisition of Second-Language Speech.* University of Amsterdam, 1990.
6. Maddieson, I. *Patterns of sounds*, Cambridge Univ. Press, 1984.
7. Möbius, B., Sproat, R., van Santen, J.P.H. & Olive, J.P. The Bell Labs German Text-to-Speech System: An Overview. *In Proc. ESCA. Eurospeech97, Rhodes, Greece*, ISSN 1018–4071, pp. 2443–2446, 1997.