Developing Speech Synthesis for Under-Resourced Languages by "Faking it" : An Experiment with Somali

Harold Somers, Gareth Evans and Zeinab Mohamed

School of Informatics, University of Manchester
PO Box 88, Manchester M60 1QD, England
{Harold.Somers, David.G.Evans}@manchester.ac.uk, zemohd@hotmail.com

Abstract

Speech synthesis or text-to-speech (TTS) systems are currently available for a number of the world's major languages, but for thousands of other, unsupported, languages no such technology is available. While awaiting the development of such technology, we propose using an existing TTS system for a major language (the base language, BL) to "fake" TTS for an unsupported language (the target language, TL). This paper describes the factors which determine the choice of a suitable BL for a given TL, and describe an experiment with a fake Somali TTS system evaluated in the real-life situation of a doctor–patient dialogue. 28 Somali participants were asked to judge the comprehensibility of 25 short Somali sentences recorded with a German TTS system. Results suggest that "faking it" provides reasonable stop-gap TTS for unsupported languages.

1. Background

Whereas currently speech synthesis or text-to-speech (TTS) systems are available for 20 or so of the world's major languages, for thousands of other, "unsupported" languages no such technology is available. While awaiting the development of such technology, we would like to try using an existing TTS system for a major language (the base language, BL) to "fake" TTS for an unsupported language (the target language, TL). Our interest is in providing support for speakers of unsupported languages in situations where their lack of ability in another language is a significant disadvantage, for example newly arrived immigrants visiting the doctor (Somers and Lovel 2003; Somers et al. 2004). TTS will play a part in spoken language translation of doctor–patient interviews, but text-based communication plays an equally important role in the pathway to healthcare, so TTS is an essential technology for users with limited or no English, and perhaps poor literacy skills in their own language. A long-term solution is to develop TTS tools for more languages, but for many the (perceived) poor ROI suggests that this is unlikely to happen soon.

The idea of faking it has been explored by Evans et al. (2002), who developed (using a slightly different techique from the one described here) fake synthesizers suitable for use with a screen reader. They report experiments with Greek, but have also used the same techniques with similar success for Albanian, Czech, Welsh, and several other languages. In a limited evaluation with just three subjects based on a number of somewhat artificial techniques standard in speech synthesis circles (modified rhyme test, nonsense syllable discrimination, tongue twisters), they claim success rates between 96% and 100%. We report a more extensive evaluation with a task simulating the real-world application of a doctor–patient consultation.

2. Faking TTS for Somali

As most readers will know, TTS systems consist of two elements: a text-to-phoneme stage, where the basic pronunciation of the text is determined, and a phoneme-to-speech stage, where the actual speech sounds are generated.

The text-to-phoneme stage involves identifying the phonemes to be uttered, and is generally done on the basis of letter-to-sound mapping rules, together with a dictionary where any irregular cases are made explicit, and some sort of syntactic analysis to disambiguate homographs and determine prosody. For fake TTS we must overcome differences in the letter-to-sound mapping rules between the BL and the TL.

In the phoneme-to-speech stage, the actual speech sounds are generated, either by concatenating prerecorded human speech or by formant synthesis.

There are three main problems for faking it: the first is choosing a BL which has as similar as possible a phoneme set as the TL. The second is overcoming differences in text-to-phoneme mapping. Finally, we must find a BL for which prosodic features such as general intonation and stress placement is as near as possible to that of the TL. Obviously, these three goals may pull in different directions, and in the end the choice of BL will involve considerable trade-offs.

After some somewhat subjective and empirical experimentation, we chose German as the BL. The main problems with this choice are phoneme set differences: six Somali consonant phonemes are not found in German: [w ɖ ɢ ʔ ħ ʕ], in orthography, respectively, *w*, *dh*, *q*, *'*, *x*, *c*. Interestingly, German glottal stop [ʔ] is not phonemic, but could be synthesized by separating two vowels with an apostrophe, as it is in Somali orthography. Two phonemes, [r] *r* and [χ] *kh*, have close equivalents in [ʁ] and [x]. In general we substituted the six "missing" sounds with German equivalents <u>, <d>, <k>, <'>, <ch>, <'>, thereby losing some phonemic contrasts: <d> served for both [ɖ] *dh* and [d] *d*, <'> for both [ʔ] *'* and [ʕ] *c*, <ch> for both [ħ] *x* and [χ] *kh*. This was one of the main causes of loss of comprehension, as our qualitative analysis (see below) indicated.

There are several differences in orthography which required us to adapt Somali spelling rules to German, for example changing *kh* to *ch*, *j* to *tsch*. As for the vowel system, Somali monophthongs, written with the vowel letters *a e i o u* correspond reasonably closely to German vowels written similarly. Long vowels can be simulated by doubling the vowel letter. Somali has five diphthongs *ay*, *aw*, *ey*, *oy*, *ow* transliterated into German as *äj*, *au*, *ei*, *äu*, *ou*. German has quite complex letter-to-phoneme

mappings, especially to do with (de)voicing, which can sometimes be overcome by doubling consonant letters. The Somali-to-"German" transliteration process is largely but not entirely automatable, so some manual revision of the texts is necessary.

A major potential problem is that Somali is a pitch accent language, with tone and stress combining to make lexical differences, e.g. *ínan* 'boy' ~ *inán* 'girl'. This is not something we can replicate, so we have to hope that context compensates for any misplaced stress tones. Otherwise, default stress placement in German seems to give a good approximation to Somali, as does the default sentence intonation. These factors are surprisingly important in the choice of BL.

Some examples of the transliterations are shown in (1)–(3).

(1) *Marka u horaysa waa in aynu samaynaa ballan.*
'First we have to make an appointment'
marka u horäisa uaa in äinu ssamäinaa ballan

(2) *Si joogto ah jimicsi ma u samaysaa?*
'Do you take any exercise?'
ssi tschoogto ah tschimi'si ma u ssamäisaa?

(3) *Miyay ku xidhantay hawada?*
'Does it depend on the weather?'
mijäi ku chirdantäi hauada?

For our experiments, we used the *RealSpeak* Solo Steffi voice from Nuance, which is a high quality concatenative sppech synthesizer. The transliterated Somali phrases were passed to the TTS synthesizer, and the outputs recorded as wav files.

3. Evaluation

Our evaluation is based on a simulation of (one side of) a patient–practitioner consultation, replacing the practitioners' questions and comments with faked Somali synthetic speech.

3.1. Problems with traditional evaluation of speech synthesis

This departs from the tradition of evaluation of speech synthesis which includes:

- the Modified Rhyme Test (MRT) (House et al. 1965, Goldstein 1995) in which subjects must match from a list of five options the word which they think they have heard;
- the Mean Opinion Score (MOS) (ITU, 1996), which involves participants in rating (some specific aspect of) output on a scale from 1 (bad) to 5 (excellent);
- Semantically Unpredictable Sentences (SUS) (Benoît et al. 1990), in which grammatical but nonsensical sentences are synthesized.

The MRT was used by Evans et al. (2002), along with a test using nonsense words, and identification of simple sentences to evaluate a faked Greek synthesizer using Spanish as the BL, but with only three subjects. We feel that these techniques, along with the MOS and SUS, both recommended by one reviewer of this paper, would be unsuitable for our Somali participants. With regard to the MOS, we have found that evaluations relying on subjective ratings are highly culture-specific. We have found that Somali participants tend not to use the intermediate points on a 7-point Likert scale in an experiment testing the relative transparency and suitability of some symbols (Johnson et al., in prep.). This supports findings reported by Flaskerud (1988), Heine et al. (2002), Lee et al. (2002), and Johnson et al. (2005). It seems unlikely to us that MOSs would give any better indication of the comprehensibility of the outputs than our current method.

Evaluation using SUS is explicitly designed to test the intelligibility of speech synthesis independent of its ability to convey meaning and its usability in a specific context. This seems to be antithetical to the aims of the current research. Furthermore, the recommended experimental method as described by Benoît et al. (1990) requires the participants to "write down what they hear on an answer sheet" (p. 388), an obvious problem if participants are largely illiterate, and/or the language in question has only loosely agreed writing/spelling conventions.

Other evaluations of speech output (e.g. Steffens and Paulus 2000) have found that overall intelligibility cannot be evaluated independently of the contributing factors, so we can assume that our measures effectively give the same information in a less abstract manner.

3.2. Simulated patient–practitioner consultation

Our methodology is based on Somers and Sugita's (2003) evaluation of SLT software in a tourist scenario. In that evaluation, the authors were interested in "the subject's ability to infer correctly the *intended* meaning of the utterance" [emphasis added] rather than the grammar or style of the translation. In a similar manner, we are interested above all in whether the faked output is intelligible, with little interest in naturalness and phonetic accuracy, unless it impinges on intelligibility, in our healthcare scenario.

In our experiment, subjects (S) were told (in their own language by a native speaker experimenter (E)) to imagine that they have gone to the clinic with respiratory difficulties, and that whatever the practitioner says was going to be translated and "spoken" by the computer. They were given examples of the computer's speech, and told what the computer was saying. They were then asked to listen to the speech samples, and to tell E what they understood. Because the syntax of the "translation" was not an issue, it was acceptable for Ss to simply repeat verbatim what they heard. E made a judgment about whether they had understood, asking clarificational questions if necessary. Ss were allowed to listen to each sample up to three times, after which they were told what the utterance was meant to be.

Five different scenarios were presented, each with a contextualisation (making an appointment, doctor asks about symptoms, doctor asks about history, nurse explains the treatment, at the pharmacist), each consisting of five phrases, giving a total of 25 items. 28 native-speaking Somali participants (19 female, 9 male) with limited or no English, aged between 17 and 55 took part in the experiments. Ss were self-selecting volunteers with some experience of attending asthma clinics with respiratory problems. Sessions lasted 20 minutes on average. Phrases were specifically tailored so as to contain essential information which had to be repeated by S if E was to judge the phrase to have been understood, e.g. 'Take this *three* times *a day*'.

At the end of each session, E elicited any general reactions and opinions from S in an informal interview.

Many of the participants commented on the poor quality of the sound: initially (for 21 Ss) we used standard low-powered plug-in PC speakers. For the final 7 Ss we provided headphones. The results for these two groups are presented separately.

3.3. Results

3.3.1. Quantitative analysis

We first consider how many samples were immediately and completely understandable on first hearing. Taking the "speakers" group of 21 Ss who each heard 25 samples, we found that 109 out of 525 samples (21%) were instantly identified correctly. For the 7 "headphone" Ss, the total of 68 out of 175 (39%) was somewhat better. At the other end of the scale, in 145 cases (28%) the sample could not be recognised even after three hearings. For the headphone group this figure is only 15 (9%).

For many of the samples, participants could get some but not all of the words at first or second hearing, so we needed to devise a scoring system that reflected this. Our scoring system is necessarily informal, but for each sample we consider how much of the significant part of the sample (key, information-bearing words) were correctly recognised after 1, 2 or 3 hearings, and try to average this. So for example, a score of 1.5 means roughly half the sample was correctly recognised at first hearing, the rest after the second hearing. A score of 3.5 means that only about half the sample was recognised, even after three hearings. A score of 2 could mean that the whole sample was recognised after two hearings, or that half of it was recognised immediately, and half after three hearings. So, the lower the score, the better.

Figure 1 shows the scores item-by-item for the participants using speakers. The "average" score is calculated based only on samples which were eventually recognised. This obviously gives a slightly inflated picture of performance, so the "adjusted average" penalises samples which were not identified even after three hearings with a score of 5. The "count" shows how many of the participants were able to recognise the sample. The figure shows very clearly that responses vary quite significantly, ranging from items 1, 21 and 22 with a maximum of 21 responses at an average of 1.33, 1.71 and 1.48 respectively, down to item 25 with only 5 responses at an average of 3.10, adjusted to 4.55.

Figure 2 shows the corresponding data for the seven participants using headphones. Response rate was 100% for 17 of the items, and 6/7 for four other items, with (raw) scores for all but two of the items comfortably above an average of 3, meaning that the items are understandable within three hearings.

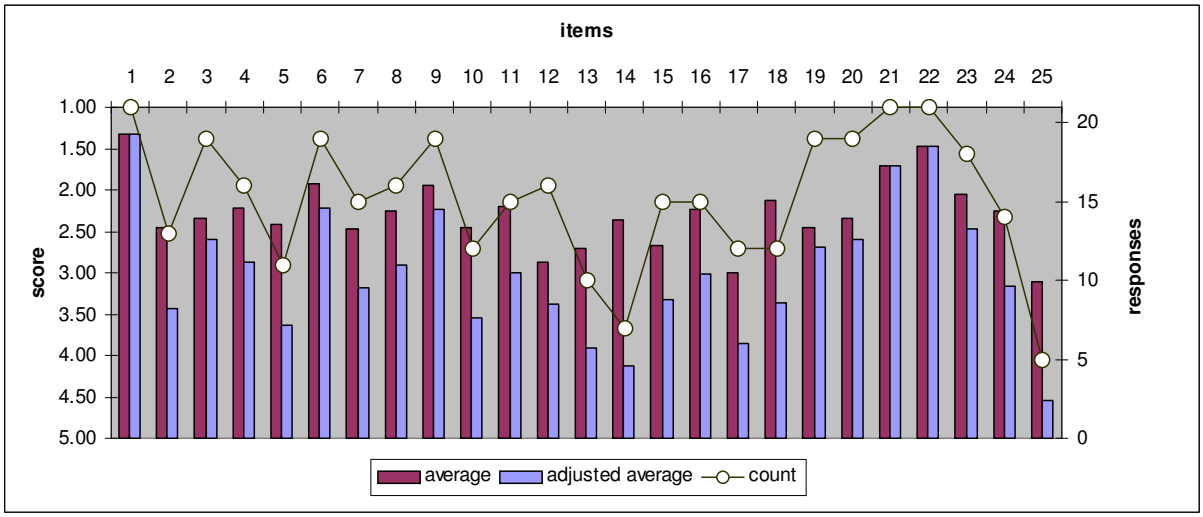

Figure 1. Overall results for participants using speakers. The scale (on the left-hand vertex) for the two averages is inverted, so that taller columns are "better".

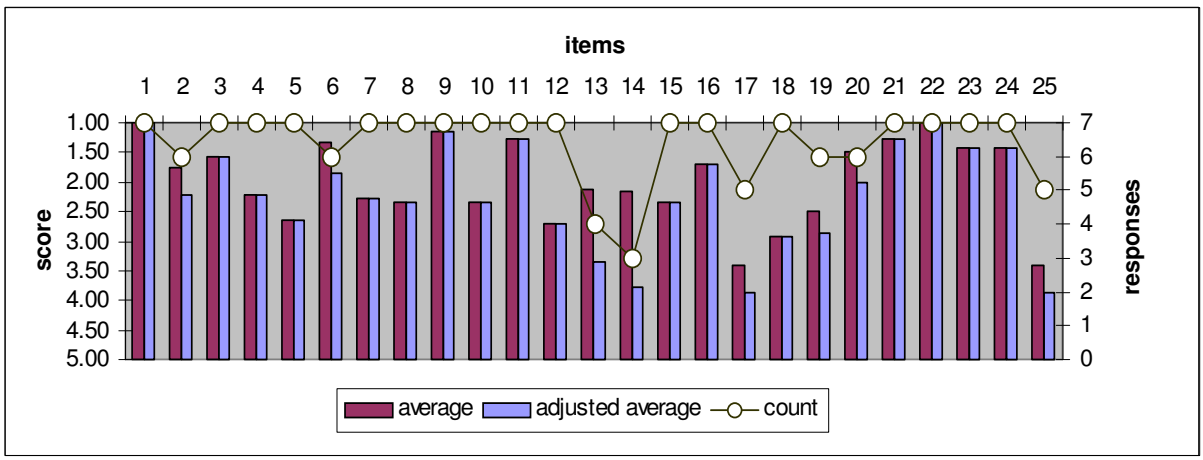

Figure 2. Corresponding results for participants using headphones.

3.3.2. Qualitative analysis

It is of interest to consider what is the source of the variation in results. Factors such as age and sex were investigated, but shown to be unrelated to performance. It might have been expected that scores would correlate with the length of the sample, but at 0.08 the correlation is barely better than random. Looking more closely at the items that scored badly, it emerges that the difficulties are linguistically based. The two main causes of failure to understand turn out to be phonetic and lexical.

Sounds that were badly rendered by the German synthesizer include *c* [ʔ] instead of [ʕ] and *x* [x] instead of [ħ]. This is the probable reason for the especially low scores for items 13 (with speakers, 10 responses out of 21 at an average of 2.70; with headphones, 4 out of 7 at an average of 2.13), 14 (7 at 2.36 and 3 at 2.17), and 25 (5 at 3.10 and 5 at 3.40), shown here as (4)–(6), with the problem words highlighted.

(4) *Ma kuu diidaa **seexashada**?*
Does it prevent you from **sleeping**?

(5) *Ma darantahay mark aad hawl **culus** qabato?*
Is it worse after **strenuous** activity?

(6) *Fadlan waa **lix** gini iyo badh.*
That will be **six** pounds fifty please.

Lexical causes of failure refer to English loan words and unusual concepts which were difficult to understand. Such words featured in items 2 (13 responses with speakers at 2.46), 5 (11 at 2.42), and 18 (12 at 2.13), shown here as (7)–(9). Interestingly, headphones alleviated the lexical problems considerably, and these cases do not stand out in Figure 2.

(7) ***Klinik**an hore ma u timi?*
Have you been to this **clinic** before?

(8) ***Warqad**da ballanta ma rabtaa?*
Do you want an **appointment card**?

(9) *Isticmaal **hayla**ha **brown**ka ah, hadii asmo sahlani ku qabato*.
Use the **brown inhaler** if you have a mild attack.

Item 17, with 12 responses at 3.00 from the speakers group, and 5 at 3.40 even with headphones) included the uncommon word *caaryo* 'mould' with the phonetically problematic *c* [ʕ] sound (10).

(10) *Musqushaada ka eeg caaryo.*
Check your bathroom for mold.

4. Conclusions

The results could be viewed in both a positive and negative light: the faked synthetic Somali is understandable, but not instantly. Some speech sounds are problematic, but in general, most of the samples were understood by most of the participants. From the practical perspective, it must be borne in mind that in reality, doctor–patient consultations with non-English speaking patients where no interpreter is present simply cannot proceed unless some support is given. In related research (Johnson et al., 2004), we have been experimenting with symbol-based communication, supported by human recordings. Simulations of asthma management consultations involve 60 or more questions and answers, involving laborious recording sessions. For a more extensive application, possibly linked to machine translation, synthetic speech must be the way forward, and while Somali remains an unsupported language, faking it may have to suffice.

References

Benoît, C., M. Grice and V. Hazan (1996). The SUS test: A method for the asessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication* 18, 381–392.

Evans, G., K. Polyzoaki and P. Blenkhorn (2002) An approach to producing new languages for talking applications for use by blind people. In K. Miesenberger, J. Klaus and W. Zagler (eds) *8th ICCHP, Computers Helping People with Special Needs*, Springer, Berlin, pp. 575–582.

Flaskerud, J.H. (1988) Is the Likert scale format culturally biased? *Nursing Research* 37, 185–186.

Goldstein, M. (1995) Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener. *Speech Communication* 16, 225–244.

Heine, S. J., Lehman, D. R., Peng, K. P., & Greenholz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference group effect. *Journal of Personality and Social Psychology* 82, 903–918.

House, A.S., C.E. Williams, M.H.L. Hecker and K.D. Kryter (1965) Articulation-testing methods: Consonant differentiation with a closed-response set. *Journal of the Acoustic Society of America* 37, 158–166.

ITU (1996) Methods for subjective determination of transmission quality. International Telecommunication Union Telecommunication Standardization Sector, Recommendation P.800.

Johnson, M.J., G. Evans, Z. Mohamed and H. Somers (in prep.) An investigation into the perception of symbols by UK-based Somalis and English-speaking nursing students using a variety of symbol assessment techniques.

Johnson, M.J., Z. Mohamed, H.J. Lovel, and H. Somers (2004) Pictographic symbols and digitised speech: a new approach to facilitating communication with non-English speaking patients. *EACH International Conference on Communication in Healthcare* Bruges, P04.02.

Johnson, T., P. Kulesa, Y. I. Cho, and S. Shavitt (2005). The relation between culture and response styles: Evidence from 19 countries. *Journal of Cross-Cultural Psychology* 36, 264–277.

Lee, J.W., P.S. Jones, Y. Mineyama and X.E. Zhang (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health* 25, 295–306.

Somers, H. and H. Lovel (2003). Computer-based support for patients with limited English. *Proc. 7th International EAMT Workshop on MT and other language technology tools*, Budapest, pp. 41–49.

Somers, H., H. Lovel, M. Johnson and Z. Mohamed (2004). Language technology for patients with limited English. *EACH International Conference on Communication in Healthcare*, Bruges, P04.09.

Somers. H. and Y. Sugita (2003). Evaluating commercial spoken language translation software. *Proc. 9th Machine Translation Summit*, New Orleans, pp. 370–377.

Steffens, J. and E. Paulus (2000). Speech synthesis quality assessment. In W. Wahlster (ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, Berlin, pp. 592–596.