Corpus Design Techniques for Irish Speech Synthesis

Amelia C. Kelly, Harald Berthelsen, Nick Campbell, Ailbhe Ní Chasaide, Christer Gobl
Phonetics and Speech Laboratory, SLSCS, Trinity College Dublin, Ireland
kellya16@tcd.ie, berthelh@tcd.ie, nick@tcd.ie, anichsid@tcd.ie, cegobl@tcd.ie

***Abstract*—Unit selection is a data-driven approach to speech synthesis that concatenates pieces of recorded speech from a large database in order to create novel sentences. Many corpora are available in the English language, including the Arctic database [1], which allows a user to create small, reliable speech synthesisers using only a small set of recorded sentences. Such resources for minority languages are scarce however, despite their increasing importance for the survival of such languages. This paper describes the current research in creating efficient Irish language corpora for speech synthesis. Corpus design techniques are discussed, in particular, two methods of data reduction that are applied to an aligned spoken corpus of Irish in order to create smaller, more efficient speech corpora.**

Index Terms: speech synthesis, corpus design, Arctic, Irish

I. Introduction

The unit selection method of speech synthesis is a data-driven, concatenative technique that draws on a large database of recorded speech, from which it can select speech segments and join them together to create novel utterances. The content of the speech database, or corpus, is vital to the performance of the synthesiser on which it is built. Naturally it is impossible to create a corpus that contains every speech sound in the language in every context in which it can be spoken, however the use of an overly large corpus will significantly slow down the performance of the synthesiser. As a result there exists a trade off between the quality and the performance of the corpus with respect to size. The Arctic database for English is an example of a corpus has been designed to address this trade-off, and is intended to be as compact as possible, while still containing the greatest diversity of linguistic units in a variety of contexts. The benefit of such a corpus is that it is freely available for download and can be used with the open-source Festival Speech Synthesis System [2] to create a personal speech synthesiser.

Despite the advances in speech technology, resources remain scarce for endangered minority languages that have experienced a decline in the number of native speakers. Irish is an example of such a language that has fallen further behind in technology development due to the lack of resources available and to the lack of commercial incentive. Since speech technology has a crucial role to play in education and accessibility, speakers of minority languages are becoming particularly disadvantaged [3].

This study aims to address the scarcity of resources for the Irish language by creating an Irish speech database that, like the Arctic database, is recorded by a number of speakers and freely available for download. This short paper outlines the approach adopted in creating an Irish language speech database and discusses two techniques[1] for designing the corpus and determining its content. The process begins with a large amount of annotated sentences, of which a subset is selected based on criteria that would deem certain sentences more suitable than others for inclusion in the final corpus. The first technique, as used to create the Arctic database, employs a greedy algorithm [4] to select only the most phonetically diverse sentences for the corpus, so that the greatest number of contextually dependent speech units are found in the smallest set of sentences. The second is a novel technique which effectively removes over-represented and therefore redundant units from the corpus, by determining which units get selected more by the unit selection synthesiser, and removing all similar ones that tend not to be chosen. Both methods result in a smaller subset of sentences being chosen for the corpus. While the first method is anticipated to select only linguistically diverse sentences, the second method will contain the best quality examples. Since the techniques will be carried out on a recorded, annotated corpus, the better method can be determined by comparing the coverage of each database and by evaluating the synthetic speech output from synthesisers based on the corpora. The techniques can then be used to reduce large amounts of data to a smaller subset before recording. The creation of small, efficient spoken Irish corpora will contribute greatly to the growing demand for minority language speech technology, in particular to the creation of Irish synthesisers.

II. Approach

Creating a corpus involves (i) selecting source material, (ii) analysing the corpus to determine unit coverage statistics, (iii) selecting the most phonetically varied sentences from the source material, and (iv) recording a speaker, [5].

A. Gathering and analysing source material

Resesarch by the Phonetics and Speech Laboratory in Trinity College Dublin has resulted in the creation of the first Irish unit selection speech synthesiser, available to use at *http://www.abair.ie*. The corpus of roughly 9,000 sentences used to create this synthetic voice draws from internet news sites and out-of-copyright works of fiction and is also used as the source text for testing the two data reduction techniques to

[1]Currently being developed by Harald Berthelsen and Amelia Kelly

create small corpus subsets. Most of the text is specific to the Gaoth Dobhair dialect of Irish, but in order to create a more adaptable speech database, a core of dialect-neutral text should be used as the basis set, and corpora of other Irish dialects can be created by adding dialect-specific texts to the basis set. The source material can then be analysed to get statistics on the frequency of occurrence of linguistic units in order to compare the linguistic unit coverage of the selected subsets.

B. Sentence selection techniques

1) The Greedy Algorithm: The greedy algorithm is an iterative technique that allows the creation of smaller corpora by choosing a subset of sentences from the basis set so that the largest number of linguistic units in context are represented in the smallest number of sentences. This is achieved by first choosing a unit size by which to define linguistic unit coverage. In this study, the base unit originally chosen was the diphone, that is the measurement from the midpoint of one phoneme to the midpoint of the adjacent one. At this time, the criteria for whether a sentence gets included in the subset depends on how phonetically varied the sentence is, given by the distribution of phones, diphones and triphones within the sentence. For example the word "cat" contains three phones (/k/, /ae/ and /t/), four diphones (the transition from silence the beginning of the word is given by /#-k/, followed by /k-ae/, /ae-t/, and the transition back to silence /t-#/) and three triphones (/#-k-ae/, /k-ae-t/ and /ae-t-#/). The algorithm will iterate through the sentences and choose the one with the most number of unique linguistic units, removing it from the basis set to be stored as the first sentence of the corpus subset. The algorithm will repeat this until a specified number of sentences have been collected. The smallest corpus that achieves maximum occurrence of these features is said to be the one with the best linguistic unit coverage.

2) The Waste Disposal Method: The waste disposal method does not focus on the linguistic variability within a sentence, but instead removes sentences from the corpus if the units in the sentence are satisfactorily represented elsewhere in the database. A sentence can be deemed redundant, and therefore removed from the basis set, if it can be synthesised using units from the rest of the corpus and the synthesised version is shown to be acoustically similar to the original recording. This can be achieved by removing a sentence from the corpus, and then attempting to synthesise that sentence using only the sentences that remain in the database. The synthesised version can be compared with the original recording using acoustic distance measures (eg. Euclidean distance between Mel-frequency cepstral coefficients (MFCC) vectors [6]) and perceptual tests (like those conducted for the Blizzard Challenge[2]). If they are similar, it may then be concluded that the sentence can be removed from the database without degrading the quality of the synthesiser.

C. Experiment Design

The data reduction techniques outlined above are performed on the basis set of 9000 Irish sentences. The greedy algorithm technique is implemented to create corpora that vary in size by increments of one hour, resulting in the creation of 8 corpora. The linguistic coverage is then ascertained for each corpus subset in order to show how the coverage increases with increasing corpus size. The most efficient corpus will be the smallest sized one that has the maximum amount of coverage. For the waste disposal method, just one corpus needs to be created that has minimum unit redundancy. The unit coverage for this corpus can then be compared with that chosen from the greedy algorithm technique. Further comparison of the methods can be carried out by recording a speaker reading the prompt sets for each corpus and creating synthesisers out of the resulting recorded speech databases. Evaluating the synthesisers by designing perception test will provide further information as to the merit of each data reduction technique.

[2]The Blizzard Challenge, http://festvox.org/blizzard/

III. Conclusion

The main focus of this research is to provide speech technology resources for the Irish language. The creation of small freely-available corpora will allow the creation of efficient and intelligible speech synthesisers, which are indispensable for use as teaching and learning resources and accessibility tools for the visually and vocally disabled. The size of the speech database used for synthesis will determine its quality and speed. In order to determine the most suitable database in terms of size and content, two data reduction techniques are described in which sentences can be selected from a large body of data to form small corpora of maximum linguistic coverage. Further comparisons can be made between the methods by evaluating the quality of synthetic voices based on the corpora. Further challenges involved in distributing the recorded speech databases are selecting and recording a speaker for each major dialect of Irish. By gathering a core set of what can essentially be considered dialect-neutral material, supplementing it with dialect-specific material, and applying the data reduction techniques described above, we hope to create freely-available Irish corpora, in keeping with the growing need for minority language speech technology resources.

IV. Acknowledgements

The CABÓGAÍ II project is funded by Foras na Gaeilge.

References

[1] Kominek, J. and Black, A. W., "CMU ARCTIC databases for speech synthesis", Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2003.

[2] Black, A. W., Taylor, P. and Caley, R., "The Festival speech synthesis system", http://festvox.org/festival, 1998.

[3] Ní Chasaide, A., Wogan, J., Ó Raghallaigh, B., Ní Bhriain, Á., Zoerner, E., Berthelsen, H. and Gobl, C., "Speech Technology for Minority Languages: the Case of Irish (Gaelic)", Interspeech, Pittsburgh, PA, 2006.

[4] van Santen, J. P. H. and Buchsbaum, A. L., "Methods for Optimal Text Selection", Eurospeech, Greece, 1997.

[5] Ni, J., Hirai, T., Kawai, H., Toda, T., Tokuda, K., Tsuzaki, M., Sakai, S., Maia, R. and Nakamura, S., "ATRECSS – ATR English Speech Corpus for Speech Synthesis", The Blizzard Challenge 2007 – Bonn, Germany, August 25, 2007.

[6] Vepa, J., King, S. and Taylor, P., "Objective Distance Measures for Spectral Discontinuities in Concatenative Speech Synthesis", Proceedings of the IEEE workshop on Speech Synthesis, 2002.