Resources Created For Building New Zealand English Voices

Catherine I. Watson[1], *Amélie Marchi*[2]

[1]University of Auckland, New Zealand, [2]Grenoble INP - ENSE3, France
c.watson@auckland.ac.nz

Abstract

This paper outlines the motivation and development of the Mansfield corpus, a new corpus to build New Zealand English (NZE) synthetic voices. In addition it describes two New Zealand English specific tools that have been developed which aid the process. First the creation and refinement of a NZE pronunciation dictionary is presented. Secondly a new web-based automatic phonetic alignment system for NZE is discussed. Finally the design of the Mansfield corpus is discussed, and some preliminary results from the first speaker recorded are presented.

Index Terms: speech synthesis, New Zealand English, speech tools

1. Introduction

There is evidence in socially interactive technologies, including robots, that the voices are important to the acceptance and impression of the interface. In [1] it was found that people approached a robot at different distances, when it employed a synthetic or natural voice. Participants went significantly closer to the robot when it had a natural voice compared to a synthesized voice. Further, those who had previously encountered an identical robot, approached the robot with the synthesized voice significantly closer than those who met the robot with the synthesized voice for the first time. Another study found jokes were perceived as funnier when spoken by a robot, rather than read from text [2]. Pucher *et al* (2008) [3] found that a local German accent synthetic voice was preferred for "fun" applications such as a talking clock or a game, but a high German accent synthetic voice was preferred for applications in the health domain.

Synthetic voice development plays an important role in the Healthbots project [6-9]. This project is dedicated to deploying robots in healthcare for the elderly. These robots have been extensively trialed at Retirement village in Auckland over a three year period [4-6], and there is a collection of robots currently being deployed at Gore Hospital. The robot currently communicates to users via a computer screen and a synthetic voice, and the users communicate via buttons on a touch screen. The robot trials [5,6] found that participants rated the robot highly in terms of overall quality of Human Robot interfaces [6], suggesting participants found the robot acceptable. There were also improvements in attitude towards the robots after meeting the robot [4,6] and the robot voice played an important part in its acceptability [7]. One of the specific aims in the project has been to give the robot a voice that reflects the population of users; currently we have developed a single New Zealand English (NZE) synthetic male voice which runs on the robot.

Tamagawa *et al* [7] showed in a New Zealand based experiment, participants rated a robots performance higher if the Healthbots robot spoke with the NZE accented synthetic voice, compared to a US accented voice. The robot was performing the same task with both voices. It was guiding participants taking their blood pressure. Participants' preference of the NZE voice over the American one in the Healthcare robot was also found by [8] in a separate study. However there has also been a strong and clear message that the existing NZE is very monotonous and needs to be improved [6-8]. The existing NZE voice [7, 9] is based on diphone synthesis, with the base speech material being over 2000 non-nonsense phrases. Whilst this method enabled very good diphone coverage, it meant the source speech material was pronounced in a monotonous manner, and this comes through strongly in the voice. The problem is, though, that building voices is a very time consuming process, our first voice took months to prepare.

We have developed a protocol to streamline and speed up the development of NZE synthetic voices. In addition to the NZE electronic lexicon that we developed for our first voice, we have now also developed a forced alignment system trained on NZE to enable automatic labelling of the speech material used to generate the synthetic speech. We have also designed and began compiling the Mansfield corpus, a purpose built corpus which will be used as prompts for the recordings for the next generation of robot voices. This paper outlines the development of these three resources, none of which have been formally presented before. The paper finishes with a brief discussion on the results to date.

2. Method

2.1 An electronic New Zealand English Pronunciation Dictionary

We developed a NZE pronunciation dictionary using the Unisyn lexicon [10,11] as the starting point. The Unisyn lexicon was designed to be able to create a lexicon for multiple English accents. The root pronunciation dictionary for the Unisyn lexicon was the Oxford Advanced Learners Dictionary, and the regional pronunciations for the different words were derived from the rules outlined in [12]. NZE is identified in [12] as sharing many features with Southern British English and Australian English. However not all the rules suggested for NZE were correct, for example the original Unisyn lexicon [11] had rules for /h/-dropping, replacing all KIT vowels with COMMA vowels (schwa), and pronouncing all "ing" endings as /In/. These were therefore not adopted for our NZE pronunciation dictionary. However the remaining the rules were adopted (see [11] for the details). These included /t/ tapping between vowels, and the /iə/ and /eə/ merger (we merged to /iə/. Although whilst for the first NZE voice we also included the GOLD phone (in words such as "hole", "role", "old"), we have since replaced all GOLD vowels with a LOT vowel in the New Zealand English lexicon. Therefore in addition to the standard 43 phonemes of NZE, the syllabics

/l!/, /n!/ and /m!/, a flapped /t/, and an unstressed /u:/ are used in the NZE pronunciation dictionary.

To this NZE lexicon derived from the Unisyn lexicon, we added new words. Firstly we added 1075 common Māori words which are found in NZE, as identified by [13]. The phones for the pronunciations of these words were taken from the New Zealand Oxford dictionary and used the NZE phoneme set, eg. "tamariki" (*children*) /ta:ma:ri:ki:/, "whanau" (*family*) (/fa:nau/). Currently the lexicon cannot cope with text with macrons, which are a feature of Maori, but not New Zealand English, therefore they are ignored. Since this is a pronunciation dictionary we have also added a series of contractions like "*I'm*", "*we're*", "*there're*" and forced the reduction of certain function words such as /əv/ for "*of*", and /ðə/ for "*the*". Finally we added a collection of words which are specific to the Healthbots project such as various drug names, and application names like "*Skype*".

Improving the NZE pronunciation lexicon is an ongoing project. The lexicon currently contains over 116 000 words, but new words are added as needs dictate. In addition as anomalies in pronunciations are being drawn to our attention, we correct them.

2.2 NZE MAUS

To assist with building New Zealand English voices we have developed a New Zealand English automatic phonetic transcription system, called NZE MAUS. This has been developed in collaboration with the Institute of Phonetics and Speech Processing in Munich and it is a new option on their MAUS system. MAUS [14, 15] is an automatic phonetic transcription system which is an open access web-based application (http://clarin.phonetik.uni-muenchen.de/BASWebServices) [16]. It has been developed for multiple languages, currently it works on 5 European languages (German, Dutch, Polish, Italian, and Hungarian), Australian English and now NZE. The pronunciation models in MAUS are based on both data-driven Markov models and pronunciation rules [15]. MAUS can be adapted to new languages mapping the phonemic inventory, formulating new pronunciation rules, and adapting the HMM to a training corpus, using a MAUS specific algorithm [15].The data for the NZE models is currently derived from the speech of 76 male and 39 female NZE speakers, predominantly from phrases, although there are some citation words. Further work is underway to improve the acoustic models, by adding a large amount of hand labelled data from 72 NZE additional speakers [17].

Each speech segment to be phonetically transcribed MAUS requires two inputs: the speech file and a text file with the words of the recording. MAUS outputs a phonetic label file, in three formats: the BAS format [15], a text grid for use in PRATT [18], or an EMU label file for use in in the EMU speech management system [19]. It is possible to process files en-masse in MAUS, making it an efficient tool for large scale phonetic annotation of data. The phonetic symbol system used in MAUS, and consequently the label files that are its output, is SAM-PA. SAM-PA is a machine-readable phonetic alphabet which maps the IPA symbols on to ASCII.

2.3 The Mansfield Speech Corpus

The end aim of the Mansfield corpus is to have a set of studio recorded single speaker databases, the primary use for these databases is for speech synthesis research. The design of this corpus was strongly influenced by CMU Artic [20,21], which was designed for building unit selection voices for use in speech synthesis. The prompts for the Artic Database were obtained from out of copyright books from the Gutenberg Project (http://www.gutenberg.org/), most of the prompts were from Jack London short stories [21]. Although the Artic prompts were available, they have been compiled for building voices with American English accents. These therefore are not so suitable for development of NZE voices. We choose to compile our own set of prompts, obtained from Katherine Mansfield's short stories (obtained from two collections: "The Garden Party and other stories", and "In a German Pension" (both out of copyright)), these were also available from the Gutenberg project. Katherine Mansfield was selected because she is a New Zealand author, and her phrasing and vocabulary has a distinctive NZE feel. Also despite being from over 70 years ago (a necessary requirement for being in the Gutenberg project), her text does have a reasonable contemporary feel to it. This is in contrast to the other literature offerings from New Zealand authors within the Gutenburg project. Out of copyright text was chosen so there would be no unforeseen complications if the robot voices are commercialized [20,21].

2.3.1 Prompt selection

Starting with an original text corpus of 8700 sentences, with 87559 words, we ended up with 1095 "nice" phrases. These were obtained using the Festvox project tools *text2utts* and *dataset_select* [20,21]. Each of the "nice" phrases was between 5-20 words long, and was easy to pronounce. The latter was determined by the density of punctuation markers, rather than due to phonetics [20,21]. All these phrases were automatically converted into strings of phones using the NZE

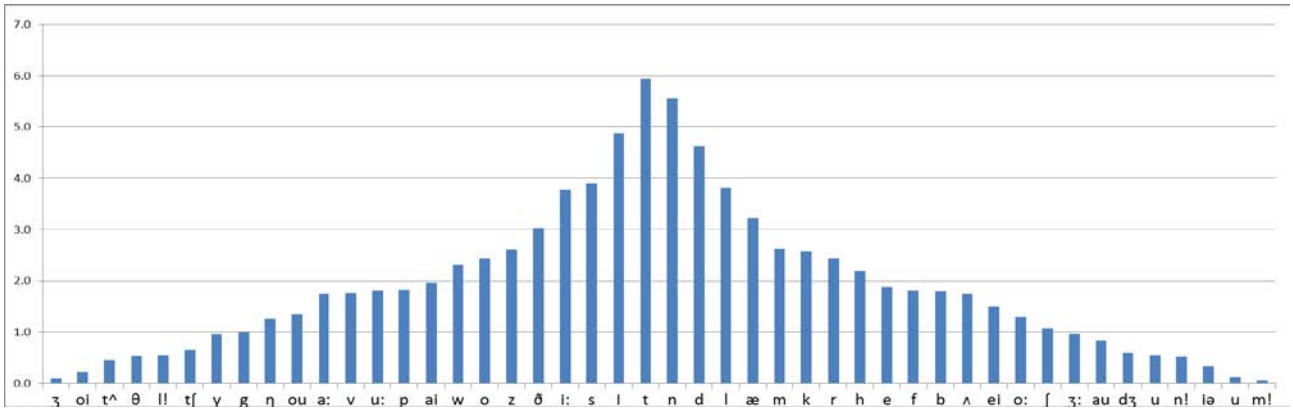

Figure 1 Frequency distribution of phones in Mansfield Corpus in % (y axis)

pronunciation lexicon, thus enabling the potential diphones to be identified. We ran the selection of "nice" phrases through *dataset_select* to ensure we had an extensive set of diphones with built in redundancy of the diphone coverage. We then, manually searched through the remaining "un-nice" portions of the text for missing diphones. These phrases were festooned with punctuation marks; which was why they were initially rejected. Through careful selection it was possible to extract out sensible phrases. Through these processes we obtained 648 prompts. Finally we added a further 35 sentences which came from other sources to increase the diphone coverage some more (see section 2.3.2 for more details on this).

2.3.2 Phone Coverage

The total number of unique phones for the NZE voice is currently 48. This comprises the 19 NZE vowels and 23 consonants, plus the three syllabics, the tapped /t/, an unstressed /u:/ vowel and silence. There are 21510 phones in the Mansfield corpus, and Figure 1 gives the phone distribution. The most common phones are /n t/, the least common phones are syllabic /m/,/ ʒ/ and the unstressed /u:/. If a diphone was found for every phone combination, with the exception of silence-silence, this yields 2303 possible phones. However not all diphone combinations are possible (eg /h/ followed by silence), and many are rarely used (eg./t/ followed by /ʒ/ or /ʊ/ followed by /ð/). With the 648 sentences we currently have 2303 diphones with the Mansfield corpus, which gives us a diphone coverage of 81.1% of all useable diphones, this compares to a 79.6% diphone coverage with the Artic corpus [22], and a 100% for the old NZE voice. Although the old voice was made up from nonsense phrases, so it was possible to get complete diphone coverage.

We did two checks to ensure the current diphone coverage would be sufficient for the existing robot function of the Healthcare robot. First we took the original robot dialogue, including all the drug names and obtained a list of all the unique diphones. All these diphones were included in the Mansfield corpus. Next we looked at unconstrained natural speech. We took the transcripts of a 13 interviews with 5 different speakers. All up there was around 31,000 words, which was about 4 hours of speech. The phone distributions were very similar to the Mansfield corpus, with the same most common, and least common phones. An analysis of the unique diphones from this set of interviews revealed we had 75 % of all possible diphones, in contrast to the 81 % coverage from the Mansfield corpus. Even so, from this set we found 31 additional diphones not accounted for the Mansfield prompt set. Phrases that contained these diphones were extracted from the transcripts and were added as prompts to the Mansfield corpus. The final prompt set contained 683 sentences

2.3.2 Recording Details

To date a single female NZE speaker using the prompt set from the Mansfield corpus has been recorded. Another speaker will be completed in the immediate future, and more are being planned. What follows is a description on this first recording, however all other recordings will follow the same process. The recordings were done in a sound-proofed Whisper room recording booth (http://www.whisperroom.com/), which is located in a large laboratory. The voice talent that was recorded sat in front of a computer screen which displayed the prompts. She spoke into a Shure SM58 microphone situated about 10-12 cm from her mouth. In addition to a speech recording, a recording of the vocal fold behavior was obtained using a Laryngograph electroglottograph (EGG) (http://www.laryngograph.com). Electrodes were placed either side of the larynx with an elastic neck band. The EGG is used to obtain accurate estimates of the fundamental frequency, which is required in the voice building process. To ensure the speech and EGG signal are time aligned the microphone is plugged into the Laryngograph microprocessor (note it was necessary to use a Tascam pre-amplifier between the microphone and EGG microprocessor to ensure the signal was at an adequate level). All recordings were done using the default sampling frequency of 16 kHz. All the recordings were passed from the Laryngograph microprocessor to a computer. This computer was outside the Whisper room, as was the computer which had all the prompts stored, and the recorder of the speech data. This ensured there was no computer noise in the speech recordings. All up it took just under 6 hours to record the speech data. The recordings took place over a number of days, as the speaker fatigued after reading prompts for over two hours. All recordings were checked, and where required re-recordings were done.

3. Results

There have been already a number of results from the compilation of the Mansfield corpus. We have developed a repeatable protocol which will enable us to record a number of different voice talents, thereby enabling us to have a collection of NZE voices which will provide the source material for the for the healthcare robots NZE voices. To date only one speaker has been recorded, but from that recording we have created the first NZE female diphone-based synthetic voice. We created this voice using the methods outlined in [22]

Table 1:The Mean and Standard Deviations for the NZE phoneme from a single female speaker.

Phoneme	Mean (msec)	Std.dev. (msec)	Phoneme	Mean (msec)	Std.dev. (msec)
i:	140.4	61.6	b	70.9	23.3
I	61.0	26.8	p	93.3	37.7
ɛ	107.8	39.5	d	72.0	29.4
æ	106.9	45.6	t	86.6	40.8
ʌ	91.4	41.4	g	81.5	31.9
a:	186.0	63.9	k	110.6	41.1
ɒ	108.2	47.4	v	54.4	22.3
ɔ:	162.7	65.7	f	113.5	39.3
ʊ	62.3	29.8	ð	54.2	17.5
u:	121.5	64.0	ɵ	89.4	38.4
ɜ:	126.9	60.5	z	116.2	56.7
u	63.6	46.0	s	144.6	48.0
ə	64.4	34.6	ʒ	95.0	23.5
ie	162.6	53.7	ʃ	124.7	38.7
ia	174.2	64.1	h	69.5	30.2
io	211.4	56.5	ʤ	107.5	36.0
ou	165.9	69.3	ʧ	133.5	45.1
au	163.2	67.9	m	77.2	34.1
i: ə	178.6	83.1	n	86.5	34.7
iə	137.0	52.3	ŋ	111.8	48.7
l!	102.2	51.5	l	75.2	48.3
m!	114.7	54.5	r	55.7	28.5
n!	129.8	49.9	w	74.0	34.9
t^	86.0	22.5	j	93.1	41.8

All synthetic voices require default duration data for the phones. We have never had duration information specific to

New Zealand English. The first NZE voice was made from nonsense phrases, where there was one nonsense phrase per diphone. As such we were unable to obtain sensible duration data from that dataset. The original NZE duration data was adapted from the duration values for the British English voice RAB in festival (http://www.cstr.ed.ac.uk/projects/festival/). The Mansfield database provides with an excellent opportunity to be able to create a duration model specific to each speaker as there are multiple exemplars for each phone. Table 1 lists the mean and standard deviation for each phone for the NZE female speaker in the Mansfield corpus.

4. Discussion and Conclusion

The design of the Mansfield corpus and protocol will enable us to streamline the development of synthetic voices. To date we are focusing on the development of unit selection voices, but we will also be investigating other approaches, such as the use of hidden markov model based speech synthesis HTS [23]. The databases within the Mansfield corpus will be very suitable for this task. To date we are not distinguishing between stressed and unstressed vowels in our diphone compilation (with the exception of the unstressed /u:/), although only the stressed vowels were used if possible. The diphones with vowels in them may be compiled in stressed or unstressed variants of the vowel (but never a schwa, unless it is specifically a diphone with a schwa in it). For future variants of the NZE voices we may investigate having separate diphones for stressed and unstressed vowels as in [21]. However at the very least we need to ensure that the diphones containing vowels are made from stressed variants of the vowels. It is easy to reduce a vowel in duration for unstressed vowels in synthetic speech, however extending an unstressed variant of a vowel is likely to result in poor quality synthetic speech.

In this paper we have outlined the development of the Mansfield corpus of NZE speech. This data comprises of speech, EEG files, and phonetic label files. To aid with this development an automatic phonetic transcription for NZE, NZE MAUS has been developed. NZE MAUS is a web-based tool and freely available for all to use. The development of this corpus of NZE would not have been possible without the NZE pronunciation lexicon, which is documented in detail here for the first time. To date there is one speaker in the Mansfield corpus, but more will be added soon. These voices will be deployed in the Healthbot robot project in New Zealand.

5. References

[1] Walters, M. L., Syrdal, D. S., Koay, K. L., Dautenhahn, K., & Te Boekhorst, R. (2008). Human approach distances to a mechanical-looking robot with different robot voice styles. In Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE Int. Symposium on (pp. 707-712). IEEE.

[2] Sjöbergh J. and Araki K (2009) Robots Make Things Funnier, New Frontiers in Artificial Intelligence: (JSAI2008) Conference and Workshops, Revised Selected Papers, p306—313

[3] Pucher, M., Schuchmann, G., & Fröhlich, P. (2009). Regionalized text-to-speech systems: Persona design and application scenarios. In Multimodal Signals: Cognitive and Algorithmic Issues (pp. 216-222). Springer Berlin Heidelberg.

[4] Broadbent, E., Tamagawa, R., Patience, A., Knock, B., Kerse, N., Day, K., & MacDonald, B. A. (2012). Attitudes towards healthcare robots in a retirement village. Australasian journal on ageing, 31(2), 115-120.

[5] Jayawardena, C., Kuo, I. H., Unger, U., Igic, A., Wong, R., Watson, C. I., & MacDonald, B. A. (2010, October). Deployment of a service robot to help older people. In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on (pp. 5990-5995). IEEE.

[6] Stafford, R. Q., Broadbent, E., Jayawardena, C., Unger, U., Kuo, I. H., Igic, A.,& MacDonald, B. A. (2010, September). Improved robot attitudes and emotions at a retirement home after meeting a robot. In RO-MAN, 2010 IEEE (pp. 82-87). IEEE.

[7] Tamagawa, R. Watson, C.I., Kuo, I.H., Macdonald, B.A., and Broadbent, E., "The Effects of Synthesized Voice Accents on User Perceptions of Robots," International Journal of Social Robots, vol. 3, no. 3, pp. 253–262, Aug. 2011.

[8] Igic, A., Watson, C.I., Macdonald, B.A, Broadbent, E. , Jayawarden, C.J, and Stafford, R., "Perception of Synthetic Speech with Emotion Modeling Delivered through a Robot Platform: An Initial Investigation with Older Listeners", in The Proceedings of the 13th Australasian International Conference on Speech Science and Technology, pp. 189–192, 2010.

[9] Watson, C., Liu, W., & MacDonald, B. The Effect of Age and Native Speaker Status on Intelligibility. NEED MORE DETAILS

[10] Fitt, S., Isard, S., 1999. Synthesis of regional English using a keyword lexicon. In: Proc. Eurospeech '99. Vol. 2. Budapest, pp. 823–826.

[11] Fitt, S. (2000). Documentation and user guide to UNISYN lexicon and post-lexical rules. University of Edinburgh, Edinburgh.

[12] Wells, J. C. (Ed.). (1982). Accents of English (Vol. 1). Cambridge University Press.

[13] Macalister, J. (2005). *A dictionary of Maori words in New Zealand English*. Oxford University Press, USA.

[14] Schiel F (1999): Automatic Phonetic Transcription of Non-Prompted Speech, Proc. of the ICPhS 1999. San Francisco, August 1999. pp. 607-610.

[15] Schiel F, Draxler Chr, Harrington J (2011): Phonemic Segmentation and Labelling using the MAUS Technique. Workshop 'New Tools and Methods for Very-Large-Scale Phonetics Research', Uni. of Pennsylvania, Jan. 28-31, 2011.

[16] Kisler, T. and Schiel, F. and Sloetjes, H. (2012). Proceedings Digital Humanities 2012, Hamburg, Germany, Signal processing via web services: the use case WebMAUS, Hamburg, pp. 30-34.

[17] Warren, P. (2002). NZSED: building and using a speech database for New Zealand English. *New Zealand English Journal*, *16*, 53.

[18] Boersma, P. & Weenink,D. (2013):Praat: doing phonetics by computer [Computer program].Version 5.3.48.

[19] Cassidy, S., & Harrington, J. (2001). Multi-level annotation in the Emu speech database management system. Speech Communication, 33(1), 61-77.

[20] Kominek, J., & Black, A. W. (2004). The CMU Arctic speech databases. In Fifth ISCA Workshop on Speech Synthesis.

[21] Kominek, J., & Black, A. (2003). The CMU ARCTIC speech databases for speech synthesis research. Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, Tech. Rep. CMULTI-03-177 http://festvox. org/cmu arctic.

[22] Black, A., & Lenzo, K. (2000). Building voices in the Festival speech synthesis system.

[23] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K. (2007, August). The HMM-based speech synthesis system (HTS) version 2.0. In Proc. of Sixth ISCA Workshop on Speech Synthesis (pp. 294-299).