Knowledge versus data in TTS: evaluation of a continuum of synthesis systems

Rosie Kay[1], *Oliver Watts*[1], *Roberto Barra Chicote*[2], *Cassie Mayo*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, UK
[2]Speech Technology Group, Universidad Politécnica de Madrid, Spain

rosiekay@gmail.com, owatts@inf.ed.ac.uk, barra@die.upm.es, catherin@inf.ed.ac.uk

Abstract

Grapheme-based models have been proposed for both ASR and TTS as a way of circumventing the lack of expert-compiled pronunciation lexicons in under-resourced languages. It is a common observation that this should work well in languages employing orthographies with a transparent letter-to-phoneme relationship, such as Spanish. Our experience has shown, however, that there is still a significant difference in intelligibility between grapheme-based systems and conventional ones for this language. This paper explores the contribution of different levels of linguistic annotation to system intelligibility, and the trade-off between those levels and the quantity of data used for training. Ten systems spaced across these two continua of knowledge and data were subjectively evaluated for intelligibility.

Index Terms: text-to-speech, speech synthesis, under-resourced languages, letter-to-sound conversion, grapheme-based acoustic modelling

1. Introduction

Grapheme-based models have been proposed for both automatic speech recognition (ASR) [1, 2, 3] and text-to-speech (TTS) [4, 5, 6] as a way of circumventing the lack of expert-compiled pronunciation lexicons in under-resourced languages.

A primary goal of our recent work [7, 8, 9, 10] has been to produce freely available tools for building statistical parametric TTS systems with little or no expert supervision. The text-processing modules we have developed are designed to construct a TTS front-end which makes as few implicit assumptions about the target language as possible, and which can be configured with minimal effort and expert knowledge to suit arbitrary new target languages. A key point of these systems is that they operate directly on a Unicode representation of the surface forms of words. It is often observed in such work that this strategy works adequately in a language such as Spanish, which employs an orthography with an approximately one-to-one letter-to-phoneme mapping.

Our experience has shown, however, that there is still a significant difference in intelligibility between grapheme-based systems and conventional ones, even in this language. Specifically, we conducted a (hitherto unpublished) experiment using data from the Albayzín 2012 TTS Challenge; this experiment was run in order to obtain a more tightly-controlled comparison of our unsupervised letter-based system [9] and a more conventional system from the challenge than had been done in the official challenge evaluation. Tighter control took the form of using identical acoustic features and voice-building recipes for both systems which mean that it was now possible to attribute the difference in performance between systems purely to the difference in system front-ends. Also, the systems were evaluated for intelligiblity (unlike in the official challenge evaluation); this was done by having human listeners transcribe speech produced by the systems and using the known text to compute word error rates (WER) of the resulting transcripts.

The difference in WERs between the unsupervised and topline systems was large (8.2% absolute: the system scores were 46.8% and 36.6% respectively) and found to be statistically significant (with α = 0.05) using the bootstrap procedure of [11]. The work presented here is motivated by these findings. This paper explores the contribution of different levels of linguistic annotation to system intelligibility, and the trade-off between those levels and the quantity of data used for training. Ten systems spaced across these two continua of knowledge and data were subjectively evaluated for intelligibility. These evaluations are designed to test 2 hypotheses: firstly, that the system using most data and most linguistic knowledge will be most intelligible. Secondly, however, we hypothesise that additional training data can compensate for a lack of linguistic knowledge.

2. Systems built

2.1. Data

The Spanish part of the Tundra corpus [12] was used for training all voices. This data is from an audiobook recording of *Don Quijote* and consists of eight hours of utterance-aligned speech in total. To deal with the variability inherent in audiobook data the lightly-supervised data selection method described in [7] was used to remove the least neutral 20% of the data. A chapter of the resulting data (119 sentences) was then held out for testing, leaving just over five hours of data for training.

To test the effect of varying amounts of training data, and the interaction of this variation with the variation in amount of linguistic knowledge given to the system, a small 1 hour set and a large 5 hour set were prepared from the training data.

2.2. Front-end annotation

Two sets of annotation were prepared for the data: one naive one based on surface orthographic forms (letters) and one based on linguistic knowledge.

For the naive annotation, the audiobook text was converted to lowercase and non-ASCII characters were substituted with ASCII-safe replacements. Whitespace was stripped and punctuation was replaced with a silence marker. Context-dependent labels were prepared from this processed text, in which each letter is characterised by the identities of the letters occurring in a five-letter window surrounding it. No further features were used (in contrast to e.g. [7] where positional and vector space model features were used).

The knowledge-based annotation was obtained from a conventional Spanish front-end which uses rule-based mod-

Level of continuum	**Knowledge**
Letters *(L)*	Five-letter context window: the system knows what the current, preceding two and following two letters are.
Phonemes *(P)*	Five-phoneme context window: the system knows what the current, preceding two and following two phonemes are.
Phonemes plus phonological class information *(P+)*	Five-phoneme context window plus additional linguistic knowledge about phonological classes, as well as place and manner of articulation e.g. vowel height, vowel length, place of articulation of consonants.
Syllable *(S)*	Syllable features as in [14]
Full *(F)*	Utterance features in [14]

Table 1: Five levels of the knowledge continuum chosen for evaluation. Each level from P+ to F makes use of knowledge at all previous levels, excluding L; letter-based knowledge is specific to the letter-based systems. Knowledge corresponds to the standard linguistic and prosodic contexts taken into account in an HTS system, as in [14].

ules to carry out tokenisation, normalisation, grapheme-to-phoneme conversion, syllabification, stress-assignment and part-of-speech tagging. It produces annotation in which phonemes are enriched with standard linguistic and prosodic contexts. [13] describes the the front-end, which has previously been used to build some of the voices in the 2010 Albayzín challenge (a competition of Spanish TTS systems).

Voices at different stages along the knowledge continuum are built by starting with a conventional feature set with all the usual linguistic and prosodic contexts accounted for, and then gradually removing this knowledge, one part at a time. There are many possible levels of such a continuum, but the 5 selected by analysis of objective evaluation of an initial set of voices built at various levels with the 1 hour data set are shown in Table 1.

2.3. Back-end construction

To train statistical parametric waveform generation modules for the systems, the speech waveforms of the training corpora were parameterised as described in [7] using the high-quality STRAIGHT vocoder [15].

For all systems, speaker-dependent acoustic models (hidden semi-Markov models) were built from this parameterised speech data and the annotation described above, using a speaker-dependent model-building recipe essentially the same as that described in [16].

3. Large-scale subjective listening test

3.1. Methodology

Table 2 lists the 10 voices subjectively evaluated. They represent all combinations of 5 levels along the knowledge continuum and 2 levels along the data continuum. The initial portion of the system identifiers indicates amount of knowledge used, as in Table 1. The final portion of the identifiers indicates the amount of data used, in minutes.

Knowledge	**1 hour**	**5 hours**
Letters	L_60	L_300
Phonemes	P_60	P_300
Articulatory information	P+_60	P+_300
Syllables	S_60	S_300
Full	F_60	F_300

Table 2: Identifiers for 10 voices evaluated in the subjective listening test.

3.1.1. Experimental design

10 voices are evaluated in the experiment, meaning each participant would hear only a small number of sentences from each voice and a very large number of listeners would be needed. This was thought to be difficult given the location of the experimenters (Edinburgh, UK), thus in order to gain good coverage, 50 listeners heard 240 sentences in two separate parts of the experiment. Sentences are from the Spanish Harvard Corpus, a phonetically balanced corpus of 720 Spanish sentences, based on the Harvard sentences[1], which have been widely used in intelligibility testing [17]. Semantically unpredictable sentences (SUS) are another common way of testing intelligibility [18], as listeners are often able to recover information in predictable sentences. However, the use of non-SUS is also well-motivated due to SUS being unrealistic in terms of the actual applications of TTS.

Each experiment followed a Latin Square design with 10 blocks of 12 sentences. The order of the blocks was randomised for each listener. Stimuli were presented through a specialised MATLAB script; the user interface had a box to type each sentence, which was heard through headphones. Sentences began to play when the user had submitted the previous response. Sentences could only be heard once, but listeners typed and moved through the experiment at their own pace.

3.1.2. Experiments

Where participants took part in both parts of the experiment consecutively they were asked to take at least a five-minute break in-between. Further, the order in which parts 1 and 2 were taken was alternated such that 50% of listeners started with part 1 and 50% with part 2. All listeners answered a short questionnaire about their variety of Spanish, age, whether they had any speech, language or hearing impairments, and how long they had been living in a non-Spanish speaking country. Participants consented to taking part, and their data being anonymised and used in the subsequent write-up, and were paid £14 for completing both parts, which lasted around 1.5 hours.

Experiments were carried out in a supervised lab with participants sitting in individual semi-sound-proofed booths, and listening to sentences through headphones.

3.2. Results

Figure 1 shows average WER for all voices, across all listeners. The letter systems perform worse than all other levels of the knowledge continuum, and there is a clear effect of training set size: all the five-hour voices (except L_300) outperform all the one-hour voices. More interesting are the relative differences between levels of the knowledge continuum: the differences appear to be more pronounced in the one-hour voices compared to

[1] http://www.cs.columbia.edu/ hgs/audio/harvard.html

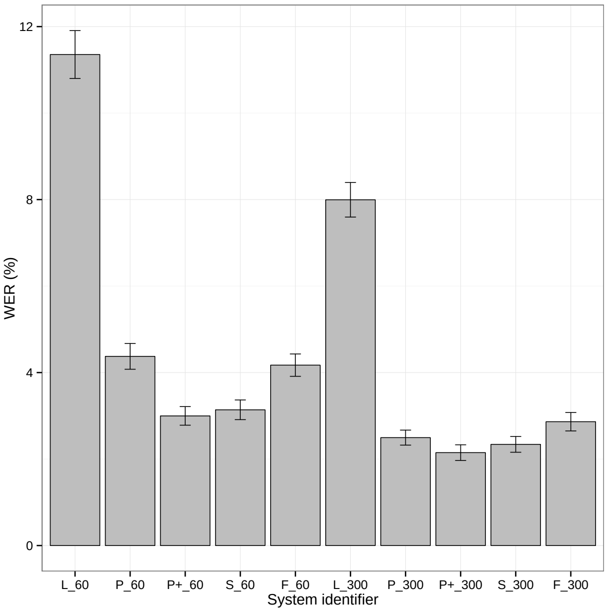

Figure 1: Average WER and standard error of all voices evaluated.

the five-hour voices. That is, additional linguistic knowledge appears to be less important when a large amount of training data is available, which supports the hypothesis that additional training data can compensate for a lack of linguistic knowledge.

Results do not support the hypothesis that F_300 would be the most intelligible. That the F voice was not the most intelligible voice is replicated at both levels of the data continuum, with tight error bars in all cases, suggesting it is not just an anomaly. In both cases, syllable, phrase and utterance level-knowledge appears to deteriorate system intelligibility. One possible explanation for this is that the decision tree clustering may have split the data too much, leading to sparsity problems. Another possibility is that there are issues with the front-end, and specifically its prediction modules for labelling the training data.

Finally, all WERs (except for the letter systems) are quite low. Due to the use of non-SUS, the sentences may have been too predictable, leading to a flooring effect. However, the sizeable gap between letter and phoneme systems suggests this cannot entirely be the case: if the sentences were entirely predictable, similarly low WERs would be expected.

Table 3 shows the results of paired Wilcoxon signed-rank tests on all voices within each level of the data continuum. This supports the hypothesis that there are greater differences between the one-hour voices than the five-hour voices, as almost all of the one-hour voices are significantly different from each other. For the five-hour voices, all voices are significantly different from the letter system, but there is only one other significant difference (between F and P+).

3.3. Post-hoc analysis: types of errors

In total, 2692 response utterances contained errors, with between one and six errors per error-containing-utterance. It was not possible to conduct a full analysis of all errors for all voices. Instead, three one-hour voices were chosen and 200 errors from each, selected at random, were examined in detail, to investigate whether the types of errors differ across voices. Explored in more detail were L_60 (the least intelligible voice), P+_60 (the best of the one-hour voices) and F_60 (the expected best one-hour voice). Three error categories were defined: replaced sound (one phoneme replaced, e.g. los → las), replaced word (more than one phoneme replaced e.g. claro → blando), and omitted word (participant made no attempt to type the word). Only three insertion errors were seen throughout the entire analysis so this is disregarded as an error category for current purposes.

	L_60	**P_60**	**P+_60**	**S_60**	**F_60**
L_60	▨	$p < .05$	$p < .05$	$p < .05$	$p < .05$
P_60	$p < .05$	▨	$p < .05$	$p < .05$	
P+_60	$p < .05$	$p < .05$	▨		$p < .05$
S_60	$p < .05$	$p < .05$		▨	$p < .05$
F_60	$p < .05$		$p < .05$	$p < .05$	▨

	L_300	**P_300**	**P+_300**	**S_300**	**F_300**
L_300	▨	$p < .05$	$p < .05$	$p < .05$	$p < .05$
P_300	$p < .05$	▨			
P+_300	$p < .05$		▨		$p < .05$
S_300	$p < .05$			▨	
F_300	$p < .05$		$p < .05$		▨

Table 3: Significant differences between systems according to paired Wilcoxon signed-rank test with Bonferroni correction for multiple comparisons

The percentage of each type of error for each voice is listed in Table 4. Most of the errors for L_60 are omissions of words altogether, rather than an incorrect guess at a word. Conversely, far fewer of the errors for the P+ and F voices are omissions: it is more likely that a word is detected, but one or more phonemes are incorrectly heard. This suggests a less serious failing on the system's part, particularly where only one phoneme is incorrect. Indeed, it seems reasonable to think of types of errors in a hierachy, with omitted word errors being of the highest magnitude, and replaced sound errors of the lowest. It is clear, then, that the letter-based voice does worse, not only in overall WER, but also in that errors tend to be more serious: when synthesis fails, it is often so bad that the listener does not attempt to guess what is heard.

One such omission comes up repeatedly for L_60: 14% of omission errors are of the word 'y' (meaning 'and'), and there are many other instances of monosyllabic words such as 'hay' 'un' and 'la' being omitted by listeners. Listening to synthesised utterances in which 'y' was omitted, there is silence where 'y' should be (see Figure 2). This strongly suggests the system failed to learn a good model for the letter 'y' (which alternates between a vowel and a consonant, depending on context: here, a vowel). This is one example where a phoneme-based system has a clear advantage, as such alternations do not have to be learned, but are already known, assuming correctly labelled data.

Voice	**Replaced sound**	**Replaced word**	**Omitted word**
L_60	15%	17.5%	67.5%
P+_60	38%	25.5%	36.5%
F_60	41%	30%	29%

Table 4: Types of error for each voice, based on samples of 200 errors selected at random per voice.

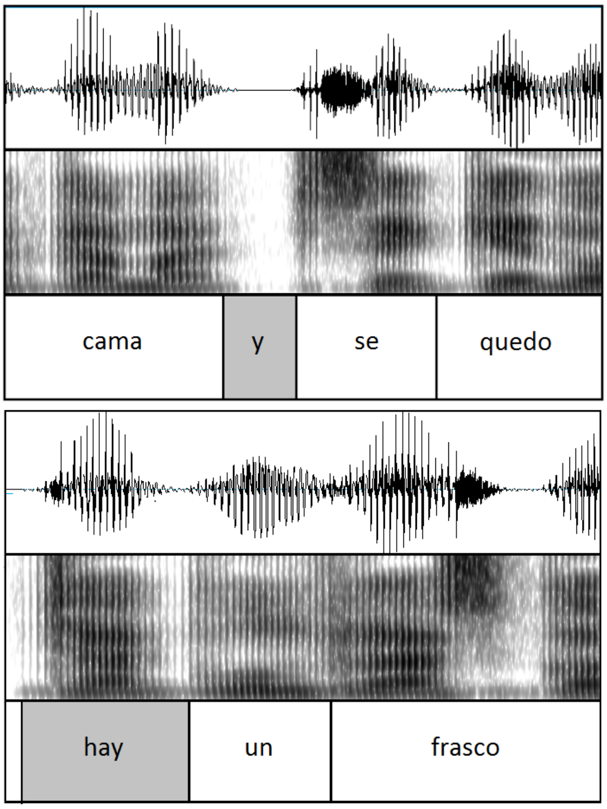

Figure 2: Spectrograms and waveforms of samples of synthetic speech from system L_60, showing a period of silence where /i/ (the word *y*) should be (top), and clear frication in the onset of the word *hay* (bottom).

'Hay' was omitted fairly frequently for L_60. Listening to utterances containing 'hay', it becomes clear that the system has not dealt with 'h' very well. 'H' in Spanish is always silent except in the letter sequence 'ch' ([tʃ]).[2] Often the system does get this right, but there are examples in which 'hay' appears to be pronounced more like 'cha' ([tʃa]) (see Figure 2): this explains the frequent omission of this word in listener responses, as 'cha' is not a Spanish word, nor does it sound close to any, leading to no attempt at transcription being made. This supports the hypothesis that two-letter to one-phoneme mappings cause problems for the system, as it is presumably learning 'h' → [tʃ] due to 'ch'→ [tʃ].

Some examples of 'l' being mispronounced with a [j] sound can also be found (presumably learned from 'll' → [j]). However, this does not seem to have caused any noticeable errors, possibly due to the phonetic similarity of [j] and [l] which causes fewer problems for listeners compared to [tʃ] vs. nothing.

A full analysis of L_300 is not made, but by listening to some of the utterances and examining some errors it is clear that whilst some alternations are still causing problems for the system some of the time, there are far fewer of the omission errors which characterise listeners' transcripts of the L_60 speech. This suggests that an increase in training data can help in learning alternations.

4. Conclusion

Our results suggest that a linguistically naive Spanish system can achieve a reasonable level of intelligibility, particularly when sufficient data is available. A WER of 5.2% for L_300 is encouraging, although perhaps could have been expected to be slightly higher, given the good letter-to-phoneme correspondence in Spanish.

Although the L systems were the least intelligible, additional training data improves the situation markedly: the performance gap between L_300 and P_300 was smaller than that of L_60 and P_60. That L_300 is much more intelligible than L_60 suggests the larger voice has done a better job at learning alternations, as many of the errors for L_60 were due to problems with these. Examining some synthesised utterances suggested these were still causing some problems for L_300, but the situation was much improved compared to the one-hour system. With five times more data, the errors are reduced by approximately a third, thus, one can tentatively consider how much data might be needed before the performance begins to match that of a phoneme-based voice, if ever.

The question of whether additional training data can compensate for a loss of linguistic knowledge certainly seems to have been confirmed: the differences between levels of the knowledge continuum were less prominent for the five-hour voices than for the one-hour voices, with fewer significant differences between different five-hour voices. This closing of performance gap when more data is used strongly suggests a larger training set can compensate for less linguistic knowledge. This is probably the most significant finding in light of the ongoing research into improving synthesis with limited resources, as training data is easier and cheaper to source than expert linguistic knowledge. It is highly encouraging that with a large enough training set, intelligible synthesis can be achieved by a linguistically naive system: this dramatically reduces the bottleneck to providing TTS in new languages. Adding low level phonological features significantly improves intelligibility over a letter-based voice, but in the absence of the necessary resources to be able to develop a phoneme set for a new language and build the necessary text-processing modules, it is promising that simply adding more hours of training data can compensate to a certain extent.

5. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement Nº 287678 (*Simple4All*).

6. References

[1] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *Proc. ICASSP*, 2002, pp. 845–848.

[2] M. Killer, S. Stüker, and T. Schultz, "Grapheme based speech recognition," in *Proc. Eurospeech*, 2003, pp. 3141–3144.

[3] Y. Sung, T. Hughes, F. Beaufays, and B. Strope, "Revisiting graphemes with increasing amounts of data."

[4] A. W. Black and A. F. Llitjós, "Unit selection without a phoneme set," in *Proc. of the IEEE TTS Workshop*, 2002, p. 7780.

[5] G. K. Anumanchipalli, K. Prahallad, and A. W. Black, "Significance of early tagged contextual graphemes in grapheme based speech synthesis and recognition systems," in *Proc. ICASSP*, 2008, pp. 4645–4648.

[2] Barring a few exceptional cases such as loan-words, e.g. hámster, or place names, such as Hong Kong.

[6] O. Watts, J. Yamagishi, and S. King, “Letter-based speech synthesis,” in *Proc. Speech Synthesis Workshop 2010*, Nara, Japan, September 2010, pp. 317–322.

[7] O. Watts, A. Stan, R. Clark, Y. Mamiya, M. Giurgiu, J. Yamagishi, and S. King, “Unsupervised and lightly-supervised learning for rapid construction of TTS systems in multiple languages from ’found’ data: evaluation and analysis,” in *8th ISCA Workshop on Speech Synthesis*, Barcelona, Spain, August 2013, pp. 121–126.

[8] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. Clark, J. Yamagishi, and S. King, “TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision,” in *Proc. Interspeech*, Lyon, France, August 2013.

[9] J. Lorenzo-Trueba, O. Watts, R. Barra-Chicote, J. Yamagishi, S. King, and J. M. Montero, “Simple4all proposals for the albayzin evaluations in speech synthesis,” in *Proc. Iberspeech 2012*, 2012.

[10] A. Suni, T. Raitio, D. Gowda, R. Karhila, M. Gibson, and O. Watts, “The Simple4All entry to the Blizzard Challenge 2014,” in *Proc. Blizzard Challenge 2014*, September 2014.

[11] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in ASR performance evaluation,” in *Proc. ICASSP*, vol. 1, 2004, pp. 409–12.

[12] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. Clark, J. Yamagishi, and S. King, “TUNDRA: A Multilingual Corpus of Found Data for TTS Research Created with Light Supervision,” in *Proc. Interspeech*, Lyon, France, August 2013.

[13] R. Barra-Chicote, “Contributions to the analysis, design and evaluation of strategies for corpus-based emotional speech synthesis,” Ph.D. dissertation, ETSIT-UPM, 2011.

[14] H. Zen, K. Tokuda, and A. W. Black, “Review: Statistical parametric speech synthesis,” *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[15] “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based {F0} extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3–4, pp. 187 – 207, 1999.

[16] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 1, pp. 325–333, Jan. 2007.

[17] C. Valentini-Botinhão, E. Godoy, Y. Stylianou, B. Sauert, S. King, and J. Yamagishi, “Improving intelligibility in noise of HMM-generated speech via noise-dependent and -independent methods.” in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[18] C. Benoit, M. Grice, and V. Hazan, “The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences,” *Speech Communication*, vol. 18, no. 4, pp. 381 – 392, 1996.