Statistical parametric speech synthesis for Ibibio

Moses Ekpenyong[a,1], Eno-Abasi Urua[b], Oliver Watts[c], Simon King[c], Junichi Yamagishi[c]

[a] *Department of Computer Science, University of Uyo, P.M.B. 1017, Uyo 520003 Uyo, Akwa Ibom State, Nigeria*
[b] *Department of Linguistics and Nigerian Languages, University of Uyo, P.M.B. 1017, Uyo 520003, Uyo, Akwa Ibom State, Nigeria*
[c] *The Centre for Speech Technology Research, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK*

Abstract

Ibibio is a Nigerian tone language, spoken in the south-east coastal region of Nigeria. Like most African languages, it is resource-limited. This presents a major challenge to conventional approaches to speech synthesis, which typically require the training of numerous predictive models of linguistic features such as the phoneme sequence (i.e., a pronunciation dictionary plus a letter-to-sound model) and prosodic structure (e.g., a phrase break predictor). This training is invariably supervised, requiring a corpus of training data labelled with the linguistic feature to be predicted. In this paper, we investigate what can be achieved in the absence of many of these expensive resources, and also with a limited amount of speech recordings. We employ a statistical parametric method, because this has been found to offer good performance even on small corpora, and because it is able to directly learn the relationship between acoustics and whatever linguistic features *are* available, potentially

Email address: mosesekpenyong@gmail.com (Moses Ekpenyong)
[1] Corresponding author

mitigating the absence of explicit representations of intermediate linguistic layers such as prosody.

We present an evaluation that compares systems that have access to varying degrees of linguistic structure. The simplest system only uses phonetic context (quinphones), and this is compared to systems with access to a richer set of context features, with or without tone marking. It is found that the use of tone marking contributes significantly to the quality of synthetic speech. Future work should therefore address the problem of tone assignment using a dictionary and the building of a prediction module for out-of-vocabulary words.
Key words: speech synthesis, Ibibio, low-resource languages, HTS

---

1. Introduction

1.1. Approaches to speech synthesis

Speech synthesis or text-to-speech (TTS) is the process of transforming a textual representation of an utterance into a speech waveform (Taylor, 2009). This transformation is generally achieved in two phases, first the *front end* transforms the text into an intermediate linguistic specification, and then *waveform generation* produces an appropriate speech signal. The first phase is necessarily language-specific and typical systems employ a great deal of language knowledge, which is embedded in resources such as a manually-written pronunciation dictionary or in the form of numerous predictive models, which have been previously learned from manually-annotated data. The second phase is less language-resource intensive, but still requires recordings of speech in the language in question. These recordings are most commonly

used on one of two ways: they are used directly, through a process of segmentation, re-ordering and concatenation to render the waveform (the *unit selection* method); or they are used to train a generative model of speech which can later be used to generate novel utterances (the *statistical parametric* method).

1.2. Speech synthesis in resource-limited situations

In this work, we use the statistical parametric method and present a comparison between several different systems which have access to a linguistic specification of varying richness. Methods which offer good performance *without* needing predictive models to explicitly predict prosody and other such "expensive" features, instead relying on "cheap" features which can be obtained reasonably directly from the text, are of particular interest in resource-limited languages. Here, we can interpret "resources" in a very broad sense: the limitations may be in terms of cost (e.g., we cannot afford to annotate training data), skills (e.g., we do not have access to people with the skills to annotate data or build predictive models), or more fundamental problems such as knowledge of the language (e.g., we do not have a definitive phoneme inventory, or a good description of the tone system, etc.).

1.3. Specific problems in resource-limited languages

Most problems lie in the front end, since this is where – in conventional systems – most of the language-specific components are to be found. A subset of these problems parallel those of text-to-speech in well-resourced languages. In text processing, we need lists of acronyms, methods for expanding number expressions, information about the possible punctuation symbols and their

functions, etc. These can be laborious to obtain for languages without large amounts of available text. Most languages require a manually-constructed pronunciation dictionary, created by skilled phoneticians. Languages without a strong written tradition may be more likely to suffer from inconsistencies in written language (e.g., spelling variation), compounding the problems of dictionary creation. The other category of problem includes those that are not generally encountered in the well-resourced case. The underlying cause in this case is usually incomplete knowledge of the language's properties (e.g., what is the phoneme inventory? is it a tone language? how many tones does it have?).

1.4. Some problematic properties of Ibibio

Although speech synthesisers for a few tone languages do exist (notably for Chinese), it is not obvious how to adapt a system from another tone language because of typological differences: Chinese has phonemic tone, whereas African languages have a broad range of morphological tone functionalities in addition to phonemic tone. For instance, in Ibibio, the subcategory of proximal/distal (temporally near or far) tense is marked by LH/HL tones on the tense morphemes. Pitch therefore has a hard mandatory semantic function rather than a soft pragmatic function in intonation languages (Gibbon *et al.*, 2006).

If tone were orthographically marked, various techniques that have been employed for other tone languages could be used to realise the marked symbolic tone acoustically. Most straightforwardly, tone-morpheme combinations could be used within a conventional state-tying system, albeit at the potential cost of a large inventory of required units or models. Possible

refinements that have been employed in systems for other tone languages include using separate decision trees for units marked with the various tones (Chomphan and Kobayashi, 2008), or a two-stage approach to the generation of F_0 designed to improve the consistency of tone realisation (Chunwijitra *et al.*, 2012). However, most such work on synthesis of tone phenomena assumes the availability to the system of correct annotation of tone. In Ibibio, however, there is no orthographic tone marking, making morphological tone assignment effectively an AI complete problem (Gibbon *et al.*, 2006) in that it requires extensive background world knowledge.

The positional dependence of tone values in the terraced tone patterning generated by automatic and non-automatic downstep in many African languages leads to a further combinatorial explosion of pitch patterning (Gibbon, 2001). In African languages, the number of inflected word forms is far larger than for languages like English or Chinese due to agglutinative inflectional morphology and complex subject-verb-object person agreement, which presents further challenges to morphological tone assignment and of course problems of data sparseness.

1.5. Prior work in speech synthesis for Ibibio

The Local Language Speech Technology Initiative (LLSTI) project provided the platform for the creation of an initial prototype adaptation procedure for a TTS system for Ibibio. A small speech database was collected, an existing Spanish text analyser was ported to Ibibio, and a waveform generation module using selection and concatenation was built. Database collection is described in (Ekpenyong *et al.*, 2008; Ekpenyong, 2009), and mentioned again in Section 3.1 in the context of the current work. This preliminary

attempt at building a voice revealed a number of previously-unconsidered problems, mainly related to front-end processing (Gibbon *et al.*, 2006). The method of porting a system between languages employed may be usable to some extent when the source and target languages are prosodically and phonemically similar, but severe problems arise when the languages are typologically dissimilar. For instance, intonation languages pose very different problems from tone languages. The synthesiser resulting from this initial work is unable to handle contextual variations in tone. The project summary presented in (Tucker and Shalanova, 2005) shows that of the languages considered by the LLSTI project, Ibibio is the most difficult language to develop TTS for, due to its complex morpho-syntactics and the fact that it is a tone language where tone is not marked orthographically.

1.6. Our approach

Developing sufficient knowledge of a language is a hard problem, probably taking decades to complete. Given this knowledge, creating the manually-annotated data resources for training predictive models is still non-trivial. An alternative approach is to attempt to use machine learning to obviate the problem of data annotation. In an ongoing line of research, we have been investigating this from a variety of angles, such as using grapheme units instead of phonemes, replacing a part-of-speech tagger with an unsupervised Vector Space Model, or relying on decision-tree based model clustering to discover patterns in the data that are the analogue of explicit manually-annotated intermediate linguistic layers – see Watts (2012) for an extended discussion. The work reported here is a simple instantiation of this approach.

The problems of tone assignment appear to be very complex indeed and

we are not claiming a solution to these. Rather, the contribution of the work presented here is to quantify the contributions to speech quality and intelligibility of different aspects of the linguistic specification, including the presence or absence of a large set of shallow positional features, and the presence or absence of tone marking.

2. The front end

Before system construction, the text of the training and test utterances was prepared as follows:

1. The removal of all punctuation except ","
2. manual correction of spelling errors in the transcript and spelling normalisation for certain words
3. loan words were transliterated to conform to Ibibio phontactics
4. manual expansion of non-standard word (NSW) tokens (e.g., numbers, acronyms)
5. manual marking of phrase boundaries

The automated text processing pipeline contained in the synthesiser front end performs the following steps:

1. tokenisation based on whitespace
2. marking of word boundaries
3. grapheme-to-phoneme conversion and syllabification using hand-crafted rules
4. simple part-of-speech (POS) tagging, using a list of known function words, with all other words being tagged as 'content word'

2.1. Phone set

We employed the Speech Assessment Phonetic Alphabet (SAMPA), a machine readable format for representing the phonemes of a language.

The phonetics of the Ibibio language are described by Essien (1990) and we represent them in Ibibio SAMPA (Gibbon *et al.*, 2006). Ibibio has a 10 vowel system (a, e, E, i, I, o, O, V, u, U), of which 6 also occur as long vowels (aa, ee, ii, oo, OO, UU), making a total of sixteen. There are 23 consonants: 6 oral stops (p, b, t, d, k, kp), 6 nasal stops (m, mN, n, J, N, Nw), 3 syllabic nasals (n, m, N), 1 trill (R\), 1 tap (r), 3 fricatives (B, f, s) and 3 approximants (j, r, R).

2.2. Pronunciation dictionary

The dictionary used was from (Gibbon *et al.*, 2004) and includes tone specifications. However, it does not mark syllabification and lexical stress. The syllabification of the utterances was therefore done automatically using rules described in (Gibbon *et al.*, 2004).

Tone information. Ibibio is tonal and has high (H), low (L), downstepped high (DH), high-low contour (HL) and low-high contour (LH) tones. In the current implementation of our system, tones are annotated manually for both the training corpus and the test utterances. Native speakers of Ibibio performed the annotation on an utterance-by-utterance basis using the audio recordings for reference. This is because the only available dictionary (Gibbon *et al.*, 2004) did not give good coverage of our data.

Positional features	Name	Description
Phoneme	p1	phoneme identity before previous phoneme
	p2	previous phoneme identity
	p3	current phoneme identity
	p4	next phoneme identity
	p5	phoneme after next phoneme identity
	p6	current phoneme identity in current syllable (forward)
	p7	current phoneme identity in current syllable (backward)
Syllable	b4	current syllable in current word (forward)
	b5	current syllable in current word (backward)
	b6	current syllable in current phrase (forward)
	b7	current syllable in current phrase (backward)
Word	e3	current word in current phrase (forward)
	e4	current word in current phrase (backward)
Phrase	h3	current phrase in utterance (forward)
	h4	current phrase in utterance (backward)
Vowel	b16	name of vowel in current syllable

Frequency features	label	Description
Phoneme	a3	in previous syllable
	b3	in current syllable
	c3	in next syllable
Syllable	d2	in previous word
	e2	in current word
	f2	in next word
	g1	in previous phrase
	h1	in current phrase
	i1	in next phrase
	j1	in utterance
Word	g2	in previous phrase
	h2	in current phrase
	i2	in next phrase
	j2	in utterance
	e5	content word before current word in current phrase
	e6	content word after current word in current phrase
	e7	from previous content word to current word
	e8	from current word to next content word
Phrase	j3	in utterance

POS features	Name	Description
Guess part of speech	d1	in previous word
	e1	in current word
	f1	in next word

Tone features	Name	Description
Tone	t1	of previous phoneme
	t2	of current phoneme
	t3	of next phoneme

Table 1: Context features

2.3. Handling linguistic context

In HMM-based synthesis, linguistic context is represented as a set of features attached to the phoneme, leading to context-depending models of phone-sized units. The full set of linguistic context features used in our system are given in Table 1. The systems compared in the evaluation use various subsets of this feature set.

This method of dealing with context is flexible, in that any level of linguistic information, from the segment up to the utterance, can easily be incorporated simply by appending a feature to the context-dependent phoneme specification. The method is potentially efficient too, because only those features that influence the acoustics should be used during model clustering (Section 3.3). In a typical text-to-speech system, several intermediate layers of information (for example, symbolic prosodic information) are used, many of them predicted from previously-predicted information. This has two disadvantages: 1) the supervised learning of these predictive models requires data labelled with the intermediate representations (e.g., ToBI symbols), which is difficult, expensive and error-prone; 2) errors are propagated through the pipeline, potentially multiplying at each stage. An alternative approach is to include only relatively shallow features – that is, features close to the text itself that can be estimated easily and robustly – and use these directly as context without attempting explicit prediction of intermediate representations.

An extreme form of this approach would be to directly use letters and not even to predict phonemes, as was done by Watts (2012). Here, we do predict phonemes, but do not attempt any explicit prediction of prosody, except for

the (currently manually annotated) phrase breaks.

Implementation. The Speect multilingual text-to-speech framework (Louw, 2008) was used to implement the front end. Speect produces a Heterogeneous Relation Graph (HRG), representing the structure (letter, word, syllable and phrase relations) of each utterance. We flatten this structure to obtain a string of context-dependent phone labels.

3. Acoustic modelling

The Hidden Markov-based Text-to-Speech (HTS) synthesis framework (Zen *et al.*, 2009a) was used to train the acoustic models and perform synthesis. The toolkit uses a statistical parametric method based on HMMs. The theory behind this approach is well-described in the literature and so we do not provide it again here; instead, please refer to Zen *et al.* (2009b) or Zen *et al.* (2007), for example.

3.1. Speech data and feature extraction

The database used here contains a total of 1,140 utterances, amounting to about two hours of speech material, read by a professional speaker. The recordings were made at the Communications Arts Studio of the University of Uyo, Nigeria, using a Marantz Professional PMD660 DAT recorder and Sony F-V420 dynamic microphone at a sampling rate of 44.1kHz. Recordings had to be conducted over several sessions because of intermittent power availability. Due to errors in the recording process, some waveforms were clipped; this was remedied as far as possible after the fact using the click, noise removal and peak restoration functions of Soundforge Pro version 10.0c.

The prompts for the speaker were sentences taken from various sources including written text (textbooks, stories, news readings and formulated sentences) and transcribed speech. The first 100 utterances were excluded from training and retained as a test set. Some examples from the test set are given in Figure 1.

The text of the sentences was determined as follows: an initial 162 sentences were selected from a corpus of transcribed news readings and few other available texts. This corpus was specially compiled for this purpose as there was no electronic text in Ibibio available. The criterion for selection was optimal diphone coverage, and this resulted in coverage of all all phonemes in our Ibibio phoneme inventory. However, initial attempts at synthesis using a concatenative voice built on this small database revealed problems in synthesising sentences containing the three rarest phonemes. Phoneme frequencies were therefore computed on the basis of these initial 162 sentences, and the rest of the 1,140 sentences in our corpus were selected and constructed to improve coverage of rare phonemes. Details can be found in Ekpenyong (2009).

Speech features. The STRAIGHT spectral analysis method was used to extract 55th order Mel-cepstral co-efficients at a 5ms frame-shift. For F0 estimation, we employed a procedure that involves voting across three different F0 trackers (instantaneous-frequency-amplitude-spectrum (IFAS), a fixed point analysis (TEMPO), and the ESPS tool "get_f0"), following Yamagishi *et al.* (2009). F0 was transformed onto the Mel scale before acoustic modelling.

ke ntak ami ammO ekemiaha akepkep akai
ekIt ebo ke ideen isaanake mbiomo ukama ufOk ifiik ibaan ikpOON
owo ndomokeet iJVNNO ibIp ibaan ibo mme eka ekON mme ke ekaa
mma kres ekpeJON akeboijoiso ke isaN ekesaNake eseet mme nsio nsio NkpO ibaan ekenaNNake eben ediwOt idVN
atie iwuot okpokoro Nka rattawu ke eboJi sted
njobio owo umOataN amaabo ke afIt nti NkpO owo anieehe oto ubOk abasi
bastO maksweed osamO amaataN OsOONO ke ufOk utom kOppa adiisVk ikakaiso iNwam
ammiiJVNNO ido ukara enie
akpodo krais ama anekke ajiire
ke mme idaha ufOk ake imO ikiide ke aNwa ufOkNwet odo

Figure 1: Sample input utterances from the test set, given in SAMPA notation

3.2. Model configuration

The modelling unit used was, as is common in statistical parametric speech synthesis, the context-dependent phone. Section 2.3 describes this approach to handling linguistic context. The amount of context used was varied in the experiments reported in Section 5, from quinphones up to "full context with tone".

3.3. Training procedure

The training utterance text is processed by the front end, resulting in a sequence of context-dependent phone labels for each utterance. These labels and the corresponding parameterised speech are used to estimate a set of acoustic models. The training process is summarized in Figure 2, showing

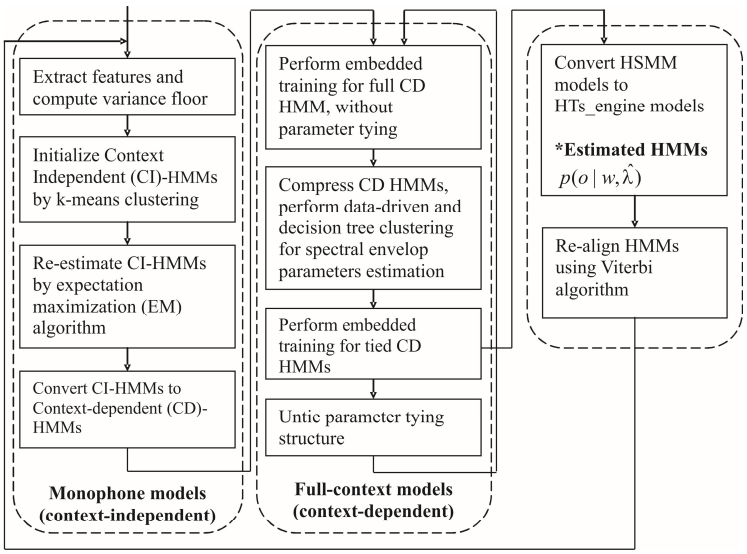

Figure 2: The training procedure

that the models are incrementally trained and refined in an iterative fashion, starting from a set of context-independent monophone models and ending up with a set of parameter-tied context-dependent models.

Decision-tree parameter clustering. A crucial step in model training is the clustering of model parameters. This is because the number of possible combinations of context feature values is very large, and only a small fraction of these will be seen in training. We used the standard decision-tree-based context clustering algorithm available in the HTS toolkit (Zen *et al.*, 2009a), which uses the minimum descriptive length (MDL) criterion (Shinoda and Watanabe, 2000). This process of clustering – in other words, finding sets of context feature values that lead to the same model parameter value – amounts to learning the relationship between the linguistic specification and the acoustic properties of the speech. By using relatively simple linguistic

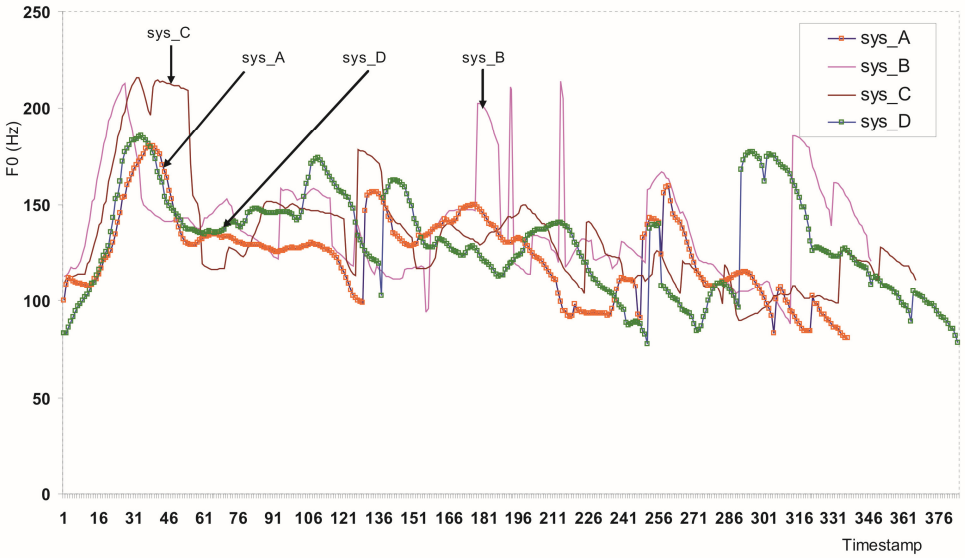

Figure 3: Examples of F0 contours for each of the 4 systems evaluated in the listening test. The Ibibio utterance (given here in SAMPA) is [eJe amaanam aNwaNa ke mme owo enie ntreubOk ke usVN OmmO keed keed] (translation in English: "S/he made it clear that people have limitations/challenges.")

features, and omitting the explicit prediction of intermediate features like prosody, the burden of predicting the acoustics rests heavily on the decision trees.

4. Synthesis

The test utterances are processed by the front end in the same was as the training utterances, resulting in a sequence of context-dependent phone labels for each utterance. From this label sequence, a corresponding sequence

of models is assembled and the waveform generation procedure is carried out. This involves two phases: first, the statistical models generate a sequence of speech features; second, this sequence is passed through the output stage of the vocoder to render the waveform.

5. System Evaluation

5.1. Dealing with tone

The aim of our experiments is to evaluate the contribution of various context factors and to understand how important each is. The prediction of tone, as discussed earlier, appears to be a very challenging problem. But there is little point in tackling that problem unless we are sure that accurate tone specification has a substantial impact on speech quality and intelligibility. This is what we will ascertain here. Therefore, all systems evaluated employ manual assignment of tone on both the training and test material (see Section 2.2 for details of tone annotation).

Before presenting quantitative results, we can compare in Figure 3 the F0 contours for each system. We observe that system A (natural speech) and system D (synthesised speech with tone labels) are relatively similar, suggesting the important of tone information.

5.2. Method

Whilst objective methods for speech synthesis quality are available, they are only useful in limited situations. Overall judgements about naturalness and intelligibility can only be reliably evaluated using subjective methods. The methods we employ are well-established and widely used, including in

System	Description	Features
A	Natural speech	-
B	Phonetic context only	p1 to p5
C	Full context, no tone	all except t1 to t3
D	Full context plus tone	all

Table 2: The systems compared in the listening test. The features referred to are explained in Table 1.

	U1	U2	U3	U4
L1	A	B	C	D
L2	B	C	D	A
L3	C	D	A	B
L4	D	A	B	C

Figure 4: The Latin Square used to assign stimuli to listening groups. Rows correspond to listener groups (L1 to L4) and columns correspond to utterance groups (U1 to U4).

the Blizzard Challenge, e.g., Podsiadlo (2007). To evaluate naturalness, we asked listeners to respond to individual utterance stimuli and respond using a 5-point scale, from which we calculated the Mean Opinion Score (MOS) for each system. Intelligibility was evaluated using two paradigms. Overall intelligibility was measured by asking listeners to transcribe Semantically Unpredictable Sentences (SUS) (Benoit *et al.*, 1996).

Three systems, listed in Table 2 were trained on the speech database described in Section 3.1. The only difference between the synthesisers (systems B, C and D) was in the waveform generation component: the context-dependent acoustic models used different subsets of the available linguistic specification. System A used natural speech recorded by the same speaker, but not used for acoustic model training.

A Latin Square design was employed to ensure that no listener heard the same utterance twice but that all listeners heard the same number of

1. <QUANT-NNPL> <DEM> e <MD> e <V> <PREP> <AJ> <NN> <NUM>. 2
2. <NN> a <COP> <AJ> <NN>. 1
3. <QUANT-NNPL> <DEM> e <MD> e <V> <NN> a <V> <PART>. 3
4. <AJ> <NN> a <V> <PREP> <QUANT-NNPL> <DEM>. 1
5. <NN> a <ASP> <V> <PART> <NN> a <MD> a <V> <QUANT-NNPL>. 3
6. <INTERROG> <CONJ> <NN> a <MD> a <V> <NN>. 2
7. <QUANT-NNPL> <AJ> <DEM> e <MD> e <V> <NN> <NUM>. 2
8. <V> <AJ> <NN> <NUM> <PREP> <NN> 0
9. <INTERJ> <V> <CONJ> <NN> <DEM> a <V> a <V> <IDEO>. 2
10. <INTERJ> <IDEO> <NN> <DEM> <V> 0
11. <PRON> a <MD> a <V> <CONJ> <NN> <PART> <ADV>. 2
12. <ADV> <NN> ibiaaNake <NN>. 1
13. <INTERROG> i <V> i <V> <NN>. 2
14. n <V> i <V> <CONJ> <INTERROG> <PART> <PRON> e <V> <NN> <NUM> <DEM> e <V>. 4
15. <NN> a_cop- <COP> <PREFIX> <V> <AJ> <NN>. 1

Figure 5: The grammar used to generate Semantically Unpredictable Sentences in Ibibio

utterances from each system, and that all systems were evaluated on the same set of utterances. Figure 4 shows the Latin Square. Twenty-eight listeners were used, partitioned into 4 groups (L1 to L4) corresponding to the rows of the Latin Square.

In the MOS test, listeners were requested to rate the utterances according to a 5 point scale where the points were labelled as follows: 1 - Not natural, 2 - Poorly natural, 3 - Fairly natural, 4 - Sounds natural, 5 - Highly natural.

5.3. Materials

Naturalness (MOS). Twenty sentences selected from a held-out portion of the corpus were used to evaluate naturalness.

Intelligibility (SUS). Utterances for the SUS test were generated by from a set of templates that we devised, inspired by the original paper (Benoit *et al.*, 1996) and shown in Figure 5. The various slots were populated randomly[2] from word lists extracted from an existing Ibibio dictionary (Urua *et al.*, 2001). Example sentences generated from the grammar are shown in SAMPA notation in Figure 6. The SUS section of the listening test involved listeners typing in the sentence they heard[3]; they were permitted to hear the utterance no more than twice.

Due to poor Internet connectivity in Nigeria and an erratic electricity supply, it was not possible to use a web-based evaluation system to ease the collation of results. We resorted to implementing the tests in Microsoft Excel.

5.4. Results and analysis

The statistical analysis and conventions for presenting the results used here closely follow those proposed by Clark *et al.* (2007) and used in all recent Blizzard Challenges.

Naturalness (MOS). A boxplot presenting the results of this section is shown in figure 7. Table 3 gives the median, mean and standard deviation of the naturalness scores for each system. Pairwise Wilcoxon signed-rank tests were used to test for significant difference between the systems. All pairs of sys-

[2]We used the same software tool as the Blizzard Challenges, kindly provided by Tim Bunnell of the University of Delaware, to do this.

[3]Or dictating it to an assistant, in the case of illiterate subjects.

mmeikua ami ekeme emianna ini mfen afIre aso itiokeet.
ideep ado ata ntok.
mmeunam ado ekId eduJeN idiONNO ademme ammo.
abVk nsiON abere ubOkubOk mmeNkukId ako.
mama a-si daap o nsO a -ma a -kVkO mme-ekON.
nsoo NkOm sabOt a-nam a-fIke mbIre.
mme-utom tVNNO ako e-ja e-fuup ifia eta.
wOOt mfa ube ikie idak sika
idIm sioNNo mma akikere ado a-kpOOn a-weeme kIraN.
ee miak nsin ami funO
mmi a-na a-tInnO utuke abiojai ake mi.
NkO akpOk i-biaNa-ke abON.
anie i-jIpO i-kpaja ekpuk.
n-faaNa i-seeRe utu mmOO ame ammO e-fan akikaak nta ako e-naa.
ajop a-do andi saN edeNNE ikiben.

Figure 6: Examples of Semantically Unpredictable Sentences in Ibibio, generated by the grammar from Figure 5 and given here in SAMPA notation, as used in the listening test to measure intelligibility.

tems (A to D) are significantly different in naturalness, except B and C, which do not differ significantly.

Intelligibility (SUS). Although the original formulation of SUS (Benoit *et al.*, 1996) suggests scoring entire sentences as correct or incorrect, it is common practice to score at the word level, taking into account insertions, deletions and substitutions (see King and Karaiskos (2009) for details), leading to Word Error Rate (WER) scores. Figure 8 presents the results as a bar chart

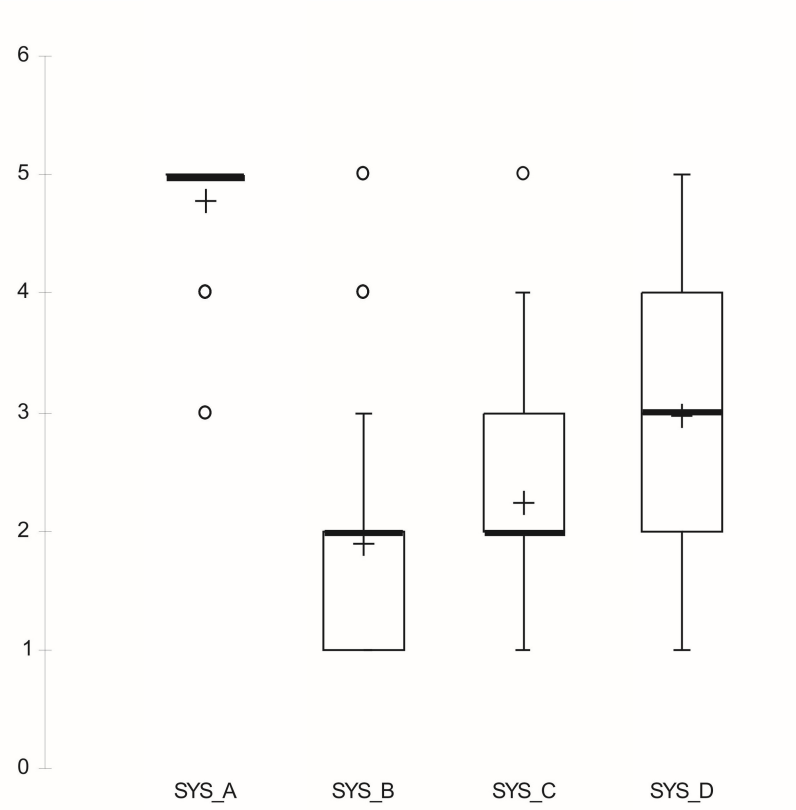

Figure 7: Boxplot showing the results of the listening test for naturalness. The median is represented by a solid bar across a box showing the quartiles, with a cross indicating the mean; whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles

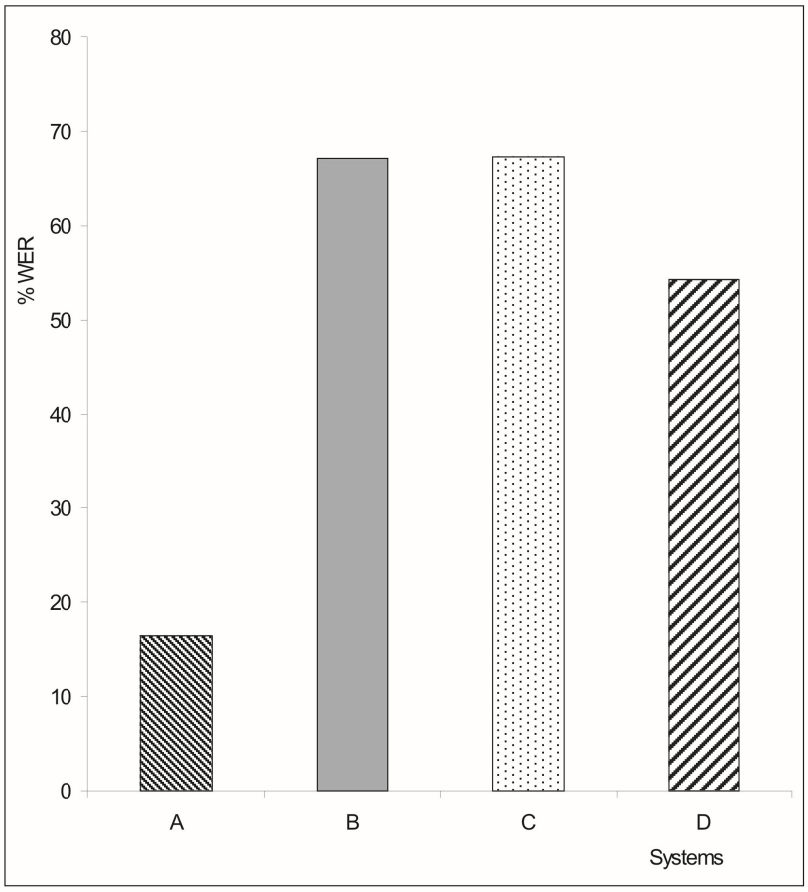

Figure 8: Bar chart showing the results of the listening test for overall intelligibility, measured as Word Error Rate on Semantically Unpredictable Sentences.

System	Median	Mean	SD
A	5	4.786	0.477
B	2	1.893	1.016
C	2	2.229	0.932
D	3	2.964	0.985

Table 3: Statistics summarising the results of the naturalness test. (SD = standard deviation)

Systems	S	I	D	C	WER
A	87	1	52	712	16.45123
B	218	2	432	322	67.07819
C	195	1	472	326	67.2709
D	214	1	287	425	54.21166

Table 4: The results of the intelligibility test using SUS. (S = substitutions; I = insertions; D = deletions; C = correct)

and Table 4 gives the scores.

6. Discussion

Whilst no system approached the naturalness of real speech (this is also always the case in Blizzard Challenge evaluations, for example), adding the tone information (system D) significantly improved perceived naturalness. Adding only the shallow context features (based on positions and frequencies – see Table 1) did not significantly improve naturalness over the naive quinphone system. Likewise, intelligibility is improved by adding the tone labels.

Word error rates (see Figure 8) are high in absolute terms, but this is because of the use of SUS, which are designed to avoid the ceiling effect that is found when using 'normal' sentences. However, the intelligibility of the synthetic speech is far worse than that of the natural speech and there

is much room for improvement here. In recent Blizzard Challenges, some synthetic systems have been found to be as intelligible as natural speech, on well-resourced languages with large databases (King and Karaiskos, 2009). This milestone should is yet to be met for Ibibio, but is not unreachable.

7. Conclusions and future work

We have presented the first statistical parametric speech synthesiser for Ibibio and evaluated three configurations of the system by varying the available linguistic context. It was found that tone specification makes a significant difference. Therefore, future work should include the use of a dictionary which marks tone and the more difficult problem of automatic prediction for novel words.

African languages present a challenge for speech synthesis. They exhibit morphological complexity which compounds the effects of a lack of resources. The lack of large datasets presents difficulties in applying either a conventional approach or a more data-driven approach. The data-driven, machine learning approach remains attractive though, since it can be applied immediately more data become available and we can reasonably expect more data to automatically lead to better quality. In the current work, a particular barrier was the lack of a large machine-readable *text* corpus for Ibibio. If this was to become available, then the unsupervised approach from Watts (2012) could be applied. In the shorter term, the use of a larger speech database is likely to give the most rapid improvements.

References

Benoit, C., Grice, M., and Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, **18**(4), 381–392.

Chomphan, S. and Kobayashi, T. (2008). Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. *Speech Communication*, **50**(5), 392 – 404.

Chunwijitra, V., Nose, T., and Kobayashi, T. (2012). A tone-modeling technique using a quantized f0 context to improve tone correctness in average-voice-based speech synthesis. *Speech Communication*, **54**(2), 245 – 255.

Clark, R., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. BLZ3-2007 (in Proc. SSW6)*.

Ekpenyong, M. (2009). Corpus design for ibibio concatenative speech synthesis system. *USEM: Journal of Languages, Linguistics and Literature*, **2**, 71 – 82.

Ekpenyong, M., Urua, E.-A., and Gibbon, D. (2008). Towards an unrestricted domain TTS system for African tone languages. *International Journal of Speech Technology*, **11**, 87–96.

Essien, O. (1990). *The grammar of the Ibibio language*. University Press Ibadan.

Gibbon, D. (2001). Finite state prosodic analysis of african corpus resources. In *Proc. EUROSPEECH 2001*, pages 83–86.

Gibbon, D., Urua, E.-A., and Ekpenyong, M. (2004). Data creation for Ibibio speech synthesis. `http://www.llsti.org`.

Gibbon, D., Urua, E.-A., and Ekpenyong, M. (2006). Problems and solutions in African tone language text-to-speech. In *International Tutorial and Research Workshop on Multilingual Speech and Language Processing*, pages 1–5.

King, S. and Karaiskos, V. (2009). The Blizzard Challenge 2009. In *Proc. Blizzard Challenge Workshop*, Edinburgh, UK.

Louw, J. A. (2008). Speect: A multilingual text-to-speech system. In *Proc. 19th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, pages 165–168, Cape Town, South Africa.

Podsiadlo, M. (2007). *Large scale speech synthesis evaluation*. Master's thesis, University of Edinburgh, Edinburgh, UK.

Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *J. Acoust. Soc. Japan (E)*, **21**, 79–86.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge University Press UK.

Tucker, R. and Shalanova, K. (2005). Supporting the creation of TTS for local language voice information systems. In *Proc. Interspeech 2005*, pages 453–456.

Urua, E.-A., Ekpenyong, M., and Gibbon, D. (2001). *Uyo Ibibio Dictionary Preprinted Draft Version 01*.

Watts, O. (2012). *Unsupervised learning for text-to-speech synthesis*. Ph.D. thesis, PhD thesis, University of Edinburgh, Edinburgh.

Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. (2009). Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing*, **17**(6), 1208–1230.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Information and Systems*, **E90-D**(1), 325–333.

Zen, H., Oura, K., Nose, T., Yamagishi, J., Sako, S., Toda, T., Masuko, T., Black, A., and Tokuda, K. (2009a). Recent development of the HMM-based speech synthesis system HTS. In *Proc. APSIPA 2009*, pages 121–130, Sapporo, Japan.

Zen, H., Tokuda, K., and Black, A. W. (2009b). Review: Statistical parametric speech synthesis. *Speech Communication*, **51**, 1039–1064.