‘MarketSpeak’ in Igbo: a speech synthesis training project

Dafydd Gibbon (Universität Bielefeld, Germany)
Ugonna Duruibe (University of Ibadan, Nigeria)
Jolanta Bachan (Adam Mickiewicz University, Poznań, Poland)

Abstract

A tutorial approach to speech technology infrastructure development for a less resourced Niger-Congo tone language (Igbo) is presented, with well-known components but a new integrative strategy for creating a prototype digital signal processing front-end for a Text-to-Speech synthesiser. Text parsing problems are only dealt with in passing because the main focus is on developing a synthetic voice for a restricted practical scenario, a market information system. We describe a generic but practical strategy for training non-specialist personnel with linguistic and/or computational skills based on a restricted domain lexicon, Finite State techniques and a traditional rule-based diphone synthesis method.

Keywords

Speech synthesis, Nigerian languages, Igbo, tone languages, resources, education.

The objective: technology training and the ‘MarketSpeak’ scenario

The objective of the present contribution is to provide a missing tutorial link in current discussions of speech technology for less resourced languages with little high quality data and incomplete descriptions, on the one hand, but with local ‘human resources’ who are not specialists in speech technology, but are trained either in linguistics or in computer science. The scientific primacy of theoretical and descriptive novelty – the ‘syntax’ and ‘semantics’ of science – is of course uncontested, but here we deal with the ‘pragmatics’ of science and technology applied to education in the field of speech technology in an environment with restricted infrastructure.

The specific task pursued in the present context is the development of components of a speech synthesiser for use with the Nigerian language Igbo (Benue-Congo, ISO 639-3 *ibo*) in the context of marketing goods and prices, the ‘MarketSpeak’ scenario, as a model for generic solutions in this field. Specifically, the present contribution presents an outcome and further development of a workshop on speech synthesis for Nigerian languages, in Abuja, Nigeria, March 2010, sponsored by a major project[1] on generic text-to-speech applications for African Tone languages.

1 Science and Technology Education Post-Basic (STEP-B), Federal Government/World Bank Intervention Project: “Towards a Generic Text-To-Speech Applications For African Tone Languages”, Grant No.: FME/STEP B/79/3/14 to University of Uyo, Akwa Ibom State, Nigeria, for Moses Ekpenyong and Eno-Abasi Urua.

The workshop goals were to provide basic training in speech synthesis for a mixed group of linguists and computer scientists who were not specialists in speech technology. Speech synthesis was chosen as a more feasible entry into speech technology than speech recognition based technologies. The specific goal of this tutorial was to create prototype 'microvoices'[2] for speech synthesis of 12 Nigerian languages by the participants, who were native speakers of these languages; this goal was achieved.

After the workshop, a synthetic voice front end for Igbo was created (Duruibe 2010), specifically for the restricted register of Igbo food markets, using straightforward traditional technology for diphone synthesis[3]. The aim was to create a first voice on which to build a rule-based Text-to-Speech (TTS) synthesiser for Igbo.[4]

The present study reports on this work, as we feel that it offers a novel approach to teaching the basics of viable speech synthesis to non-specialists. For this purpose we introduced new heuristic procedures for automatically creating a phonetically rich data set for recording, for automatically extracting diphones from speech data, and for evaluating data quality and system quality by providing close copy gold standard benchmarks. The new contribution of the present study lies not in the development of novel theories, models, algorithms and application domains (Roux et al. 2010) but in novel combination and deployment of known technologies in new fields of application for less resourced languages, for non-specialists with basic linguistic and/or computing knowledge. In this context, teaching strategies and low budget speech synthesis development methods must be used, with the focus on a community with a language which so far has few empirical, descriptive and technological resources but a strongly felt need for and interest in technological development.

An optimal solution for the task, if it were solely system development and not an educational task, would be an easy-to-use speech synthesis kit with clear linguistic interfaces and user-friendly tools for data selection, processing and evaluation, but unfortunately, to date there is no such kit. Consequently, criteria for using results from different areas of linguistics, computational linguistics and speech technology were integrated for this purpose. The present contribution concentrates on the requirements and creation of the phonetic and digital signal processing (DSP) components of a text-to-speech synthesiser. The Natural Language Processing (NLP) components and their computational linguistic foundations are only dealt with in passing.

---

2 A 'microvoice' is defined as a speech synthesis voice created from a restricted data set without the aim of modelling the entire sound system of a language. Microvoices are typically used in tutorial contexts and in phonetic research contexts for creating synthetic speech for experiments.

3 The well-known MBROLA 'legacy' diphone synthesis front end (http://tcts.fpms.ac.be/synthesis/) was chosen, for reasons which will be explained in the text.

4 In the context of a Festschrift contribution for Justus Roux, with whom the first author has had many inspiring discussions in many environments, it is appropriate to combine Justus' interests in speech technology, particularly speech synthesis, African languages, tone and technology infrastructure development in a cooperative advisor-student paper.

In Section 2, software requirements and design are discussed, followed by discussion of linguistic specifications of the system prototype in Section 3. In Section 4 the pre-recording, recording and post-recording phases of data processing for voice development is dealt with in some detail, and the workflow is presented. Section 5 briefly describes the voice construction step, and Section 6 presents conclusions and an outline of future work.

Requirements

Software selection requirements

The general requirements for providing a feasible method for rapid speech synthesiser development for Nigerian tone languages have already been outlined. The general requirements lead to a number of specific requirements which determined the choice of the speech synthesis method. The choice was a conservative one, which fell on the MBROLA diphone synthesiser, for the following reasons:

1. Suitability for use in a training context with minimal training time for linguists whose knowledge of computation is limited to using consumer software, and computer scientists with no more than a basic knowledge of phonetics.
2. Free software, because of minimal or no funding.
3. Comprehensive documentation (Dutoit, Pagel 1996; Dutoit 1997[5]), to facilitate understanding of procedures and to encourage further creative development.
4. Credentials of extensive use for multilingual speech synthesis (73 voices for 36 languages are publicly available on the internet).
5. Cross-platform availability (runtime binaries for 37 operating systems and operating system versions are available, including the required Linux and Windows versions).
6. Offline use, independence from internet tools, because of the expense and unpredictability of internet acccess.
7. Ease of installation, to facilitate deployment by non-specialists.
8. Simple interface between text parser and diphone synthesis components, to permit close cooperation between linguists, computational linguists and phoneticians.
9. Reasonable quality in relation to the other requirements.

Clearly MBROLA is not a state of the art system any longer, but there is no other speech synthesis system which fulfils requirements 1-8 above to anything approaching the extent to which they are fulfilled by MBROLA, though there are better quality synthesisers. Although 36 languages are represented in the public MBROLA voice collection, the only African language represented there is Afrikaans, which is historically and typologically unique on the continent in being closely

5 Cf. also the MBROLA web page: http://tcts.fpms.ac.be/synthesis/

related to European languages. Thus there is a lot of room for further development and experimentation in relation to typologically different African languages.

System architecture requirements

Figure 1 shows the overall architecture within which the speech synthesis front-end for Igbo is designed. Currently the focus is on the diphone synthesiser DSP front-end, i.e. the component which produces acoustic output from a linguistically oriented representation of pronunciation, containing both segmental and prosodic information. The computational linguistic foundations of the back-end NLP components of preprocessing, parsing and the automatic creation of pronunciation models were not created in the prototype, and for this reason the NLP back-end is greyed out in Figure 1, though some aspects are briefly outlined in Section 3, in the discussion of linguistic specifications for Finite State grammar and tone modelling.

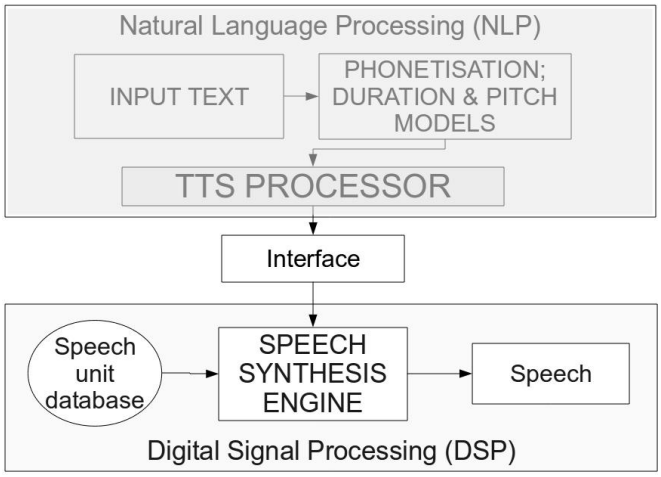

Figure 1: General architecture model for text-to-speech synthesis.

The important feature of the MBROLA concept, which is not always found in other speech synthesis concepts, is the *Interface* specification, which permits both segmental units and their prosodic properties to be specified in a linguistically transparent fashion. The interface is implemented as a text file with the extension ".pho" (informally referred to as a 'pho file'. The structure of the pho file representation is defined as follows:

<pho-file> ::= <pho-line>*
<pho-line> ::= <commentline> | <phonemespec> | emptyline
<phonemespec> ::= phoneme duration <pitchspec>*
<pitchspec> ::= position hertz
<commentline> ::= # char*

In other words, the pho file representation consists of an arbitrary number of lines (actually depending on the length of the utterance to be synthesised), which may be either a comment starting with a hash character "#", a phoneme specification, or an empty line.

The phoneme specification contains three kinds of information:

1. A *phoneme label*, usually in the keyboard-friendly SAMPA encoding of the International Phonetic Alphabet (cf. Gibbon et al. 2000a:359ff., also available via internet search). Sometimes major allophones are specified in addition to phonemes.
2. A *duration value in milliseconds*, specifying the length of the phoneme in the context of its preceding and following neighbours. The duration value is in general determined by statistical analysis of durations in a corpus of annotated speech, often by means of a classification and regression tree (CART) which weights different factors found in the corpus.
3. A *pitch contour specification*, consisting of a series of zero or more *pitch value specifications*. Each pitch value specification consists of a *position specification* as a percentage of the phoneme/phone (early, e.g. 20%, mid, i.e. 50%, late, e.g. 80%) and the *pitch specification* in Hertz. By supplying a sequence of specifications, tonal contours can be emulated. Voiceless sounds and pauses are generally not supplied with a pitch value specification.

The interface content specification determines not only the data requirements for the recording phase of system development but also the requirements for phonetic information:

1. The phoneme inventory of the language concerned (perhaps with major allophones).
2. A duration model for phoneme lengths.
3. A pitch model for specifying the shapes of contours on specific phonemes.

An example of an interface specification for Igbo is shown in Table 1.

Table 1: Input interface table to MBROLA diphone synthesiser front-end.

Phoneme	**Duration**	**Pitch specification**					
(SAMPA)	**(ms)**	**%pos**	**Hz**	**%pos**	**Hz**	**...**	**...**
_	200						
a	301	80	180				
gw	223						
a	169	80	145	(only one pitch specification per line in this file)			
_	1445						
O	331	80	234				
k	162						
a	231	80	145				
_	1296						

System workflow requirements

The system architecture is centred on the interface between the NLP and the DSP components, and on the construction of phoneme models with their duration and pitch characteristics. The phoneme model is taken from linguistic analyses of Igbo (see below), and the duration and pitch

characteristics are taken from computational corpus analysis of the recordings, implemented with Python and Unix/Linux shell scripts. The overall workflow which is required for producing the Igbo voice is derived from the architecture, and is shown in Figure 2. The figure is intended to be self-explanatory, given the previous discussion. However, for clarification the main inputs and workflow phases are described more fully in context.

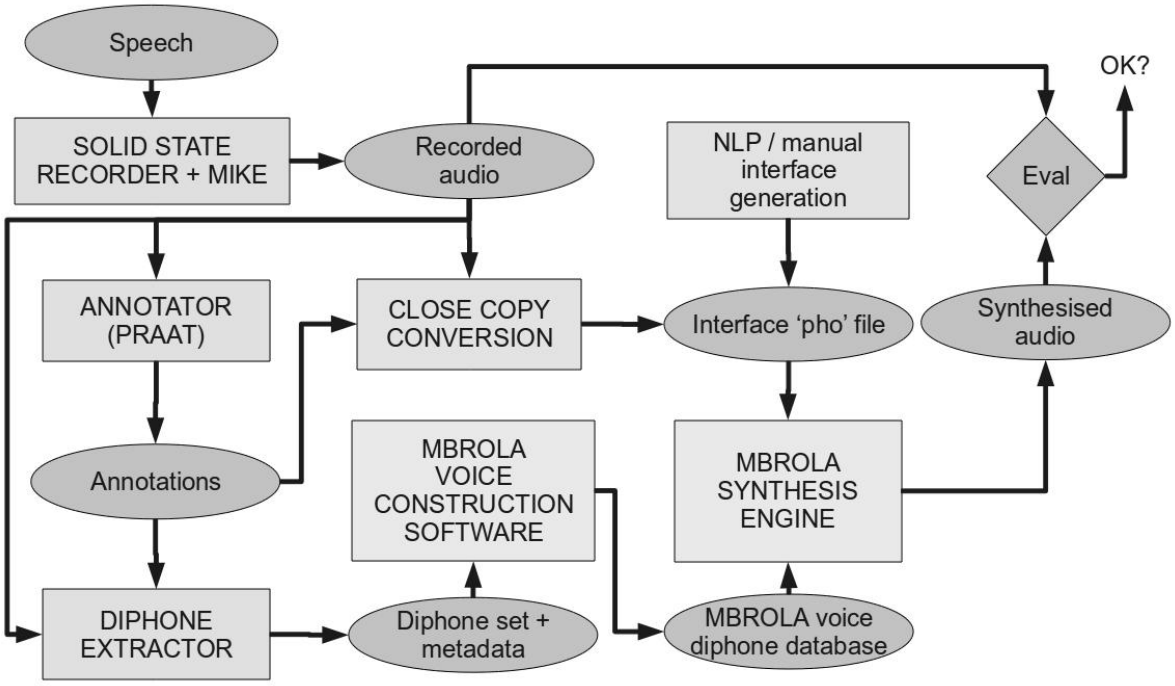

Figure 2: Workflow for Igbo diphone voice development.

First, speech input for voice development is digitally recorded, and used for 3 purposes:

1. Evaluation – the ultimate 'gold standard' for comparison with the voice output.
2. As the basis for the (manual) creation of annotation files, e.g. with the Praat toolbox ():
    1. for use in the automatic (or manual) creation of close-copy synthesis input with the voice, for 'gold standard' evaluation. For this purpose, the phonemes and their durations are extracted from the annotation, and the pitch values are extracted from the speech signal, and formatted in pho file interface format;
    2. for use, together with the speech signal, in automatically extracting diphones from the speech signal by means of a Python script, as input for the MBROLA voice creation procedure.

Second, the diphone set is created from the corpus. Using Python scripts, the metadata about the diphone timestamps in the speech database are extracted from the speech annotations, and diphone files are extracted from the speech file for input to the MBROLA voice construction software (the 'Mbrolator'), which processes the diphone files into an MBROLA voice.

Third, the inputs to the MBROLA runtime synthesis engine are a pho file interface (created manually, or automatically with an NLP component), and the MBROLA voice. The output of the synthesiser is evaluated in perception tests, in which it is compared with utterances from the original speech database.

Finally, the voice is evaluated in perceptual tests using the pho file representations based on close copies of the original speech and on automatically generated versions.

Linguistic specifications

Grammar component: nominal and verbal sequences

For the purpose of modelling grammar in the present scenario a Regular (Type III) Grammar or Finite State Automaton (FSA) model is assumed to be adequate. The conditions for which a general context-free model is needed, i.e. centre-recursive sentence embedding of expressions, are more typical of formal written text, and for practical purposes can be excluded from models of restricted spontaneous speech.

Examples of finite state models for typical nominal and verbal sequences found in Benue-Congo languages like Igbo are shown in Figure 3 (Gibbon et al. 2003). The models were originally developed for the neighbouring and both historically and typologically closely related Benue-Congo language Ibibio (ISO 639-3: ibb).

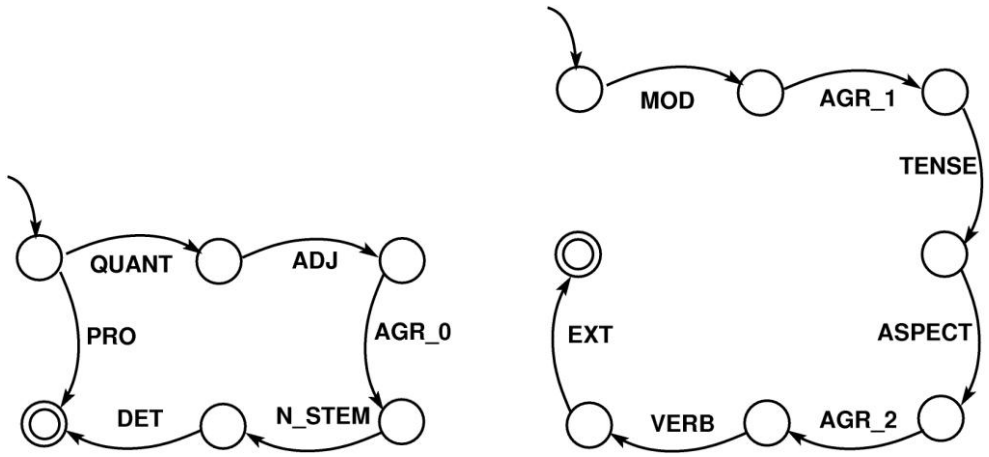

Figure 3: Ibibio Noun Phrase (left) and Verb (right) sequences.

Figure 3 shows two transition diagrammes representing one Finite State Automaton (FSA) for Noun Phrases and one for Verbs. Transition diagrammes can be seen as a kind of map, in which the states (represented by circles) symbolise places and the transitions (represented by arrows) represent events involving motion from one place to another (e.g. recognition or generation of a word). The starting point is represented by an arrow with no circle attached to the beginning, and terminal

states are represented by a double circle. Classes of words, rather than individual words, are shown, labelled with grammatical categories.

The Noun Phrase transition diagramme shows that a Noun Phrase can be either a PROnoun, or a sequence of QUANTifier, ADJective (optionality not shown), a Subject AGReement morpheme (AGR_0), a Noun stem and a DETerminer morpheme. In addition to lexical tone, each of these elements is modified by a tonal specification.

The Verbs are agglutinative: a stem concatenated with sequences of prefixes and suffixes. The Verb transition diagramme shows that Verbs consist of a MODality morpheme, an AGReement morpheme (AGR_1) which agrees with the Subject AGReement morpheme (AGR_0), a TENSE morpheme, an ASPECT morpheme, an Object AGReement morpheme which agrees with the AGR_0 morpheme of the Object Noun Phrase, and an EXTension with various functions. In addition to lexical tone, each of these elements is modified by a grammatical tone specification, and tone agreement is subject to further constraints.

Initial investigation shows that Igbo and Ibibio grammar do not differ greatly in these respects, though the vocabulary and detailed constraints certainly differ. When an NLP component is added to the present voice prototype, it will be based on a grammar of this type.

Tone sequences

The tone sequencing properties of Igbo involve tone terracing, a special case of the downtrends outlined by Connell (2002). The fundamental frequency (F0) values associated with each individual phoneme are not lexical properties of the phoneme but dependent on several factors, expressed by a function with at least the following six factors:

$$F0(phoneme_i) = f(baseline, onsetpitch, declination^i, perturbation, tone, intonation)$$

The index *i* indicates the position of the phoneme in the sequence. The *baseline* factor is a value below which the frequency does not fall. The *onsetpitch* factor is the initial pitch minus the baseline. The $declination^i$ factor is generally <1 and the power superscript determines an asymptotically downtrending frequency trajectory, depending on intonation factors, e.g. in questions this factor can also be >1, or have an additive pitch-raising element. The *perturbation* factor is the effect of the modification or blocking effect of the consonantal or vocalic segment $phoneme_i$ on the frequency. The *tone* factor is the lexical or grammatically determined contrastive tone (cf. Gibbon et al. 2009; this factor corresponds structurally to *accent* in a stress/accent language). The *intonation* factor is a complex global pattern associated with speech acts and focus (Hirst et al. 1998; for Igbo cf. Ikekeonwu 1993). The *baseline*, *onsetpitch* and $declination^i$ factors together determine the overall terracing downtrend. For related work, cf. Liberman et al. (1993), Pierrehumbert and Liberman (1994), Akinlabi and Liberman (2000).

The finite state model which accounts for the basic tone terracing pattern for Niger-Congo languages is shown in Figure 4. The simple model applies in principle to Igbo, but does not apply equally to all Niger-Congo languages (even apart from toneless cases such as Swahili): the case of Baule (Kwa, ISO 639-3 bci) is an example of a case in which added complexity is required in order to account for three-tone sequence effects rather than effects between two neighbouring tones (Gibbon 1987; cf. also Gibbon 2001, 2009 for further details).

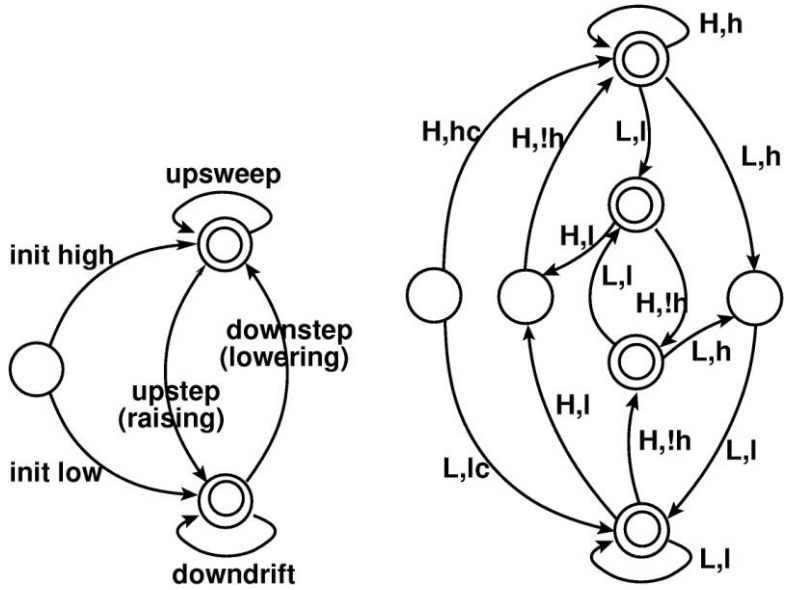

Figure 4: Tone sequence models: general tonological model for Niger-Congo 2-tone languages, and specific model (with lookahead) for Baule.

In the phonetic interpretation of the categories represented in the transition diagrammes of Figure 4, patterns such as downstep, downdrift, upstep and upsweep are modelled by operations on the pitch of the immediately preceding Tone Bearing Unit (TBU), relative to a reference pitch baseline. The modifications are logarithmic, implemented by multiplication of the preceding pitch value by a factor < 1 or > 1, depending on High or Low tone, with a cumulative logarithmic effect in the course of the utterance. In principle, the procedure resembles that described for Igbo by Liberman & al. (1993) and fomulated by Akinlabi & Liberman (2000) as follows:

1. Divide the utterance into maximal regions of like tone [represented in the automata by the local loops - Authors].

2. Place a mid-valued tonal target at the start of the utterance [the ‘init_low’ and ‘init_high’ events - Authors].

3. Place a tonal target at the end of each region, chosing an F0 value determined by the tonal type, downdrift/downstep, and final boundary effects if any [not necessary in the automata - Authors].

4. Interpolate linearly from target to target [the events on the loop transitions - Authors].

However, the particular non-local ‘lookahead’ strategy selected by Akinlabi and Liberman is not confirmed in phonetic investigations on sequences of like tones in neighbouring Ibibio (Gibbon & al. 2000b), where the logarithmic approach is found, and better represented by the local ‘immediate neighbour’ strategy of the automata and their phonetic interpretation. It is an empirical question whether Igbo differs so much from Ibibio in this respect.

Data

Data and data processing requirements specification

The data and data processing requirements are derived from the task and architecture specifications and may be summarised as follows:

1. Pre-recording phase: the scenario vocabulary, the scenario prompts for recording, definition of the phoneme inventory, definition of the diphone inventory.
2. Post-recording phase: annotation of recordings; extraction of diphone time-stamps; extraction of diphone files; assignment of phonemes, durations and pitch specifications in pho files.

The scenario prompts (wordlist and sentence list)

The prompt list comprises words from the semantic fields *Food Items, Domestic Animals, Fruits and Vegetables*, *Kitchen Utensils*, *Other Items*, and *Market Days*, as well as assorted sentences, including formulaic greetings. To illustrate the lexical data, the word field *Domestic Animals* is shown in Table 2.

Table 2: Sample Igbo vocabulary database table from semantic field ‘Domestic Animals’.

Orthography	IPA transcription	SAMPA transcription	Corpus Tones	Gloss
ehi	/ehi/	ehi	H-H	‘cow’
ewu	/ewu/	ewu	H-H	‘goat’
atụrụ̄	/atʊrʊ̄/	atUrU	H-H-!H	‘lamb’
ọkụkọ̀	/ɔkʊkɔ̀/	OkUkO	L-H-L	‘fowl’
ezì	/ezì/	ezi	H-L	‘pig’
anụ	/anʊ/	anU	H-H	‘meat’
ejule	/edʒule/	edZule	H-L-L	‘snail’

Phoneme inventory

The phoneme inventory of Igbo is known (cf. Duruibe 2010) and does not need to be specified phonetically here. More important is the presence of an adequate number of phoneme instances. A phoneme frequency list for the ‘MarketSpeak’ corpus is shown in Table 3. Including pause, Igbo has 38 phonemes according to the analysis used here for Igbo voice construction: /a/, /b/, /ʧ/, /d/, /e/, /ɛ/, /f/, /g/, /ɓ/, /ɣ/, /gʷ/, /h/, /i/, /ɪ/, /ʤ/, /k/, /kʷ/, /l/, /m/, /n/, /ɴ/, /ɲ/, /ŋʷ/, /o/, /ɔ/, /p/, /ƥ/, /r/, /s/, /ʃ/, /t/, /u/, /ʊ/, /v/, /w/, /j/, /z/, plus pause. The phonemes are represented for computational purposes with

the X-SAMPA IPA encoding for keyboarding convenience (cf. Gibbon et al. 2000a:359ff.); the pause is represented by the understroke "_".

For the present contribution, justification of the details of the descriptive phonetic definitions of specific phoneme labels is less significant than having a complete inventory of phonemes together with accurate labelling of the recorded speech signal, because the phonemes are mapped in diphone contexts directly to the acoustic representation obtained from the corpus, and not to a phonetic representation of the traditional symbolic type, thus capturing local allophonic variation.

Interestingly, this direct mapping from phonemes to acoustic representations corresponds exactly to the view of Bloomfield (1933), who did not consider the intermediate level of phonetic representation in symbols to be particularly important for the development of a realistic theory.

Table 3: Corpus frequency list for 'MarketSpeak' Igbo microvoice.

112	_	24	n	11	z	5	s	3	j
76	a	24	E	10	kw	4	p	2	NX
39	U	23	o	9	tS	4	Nw	1	w
37	O	21	k	9	dZ	4	J	1	v
37	I	19	r	7	h	4	gw	1	S
29	e	17	u	7	b	4	f	1	G
28	i	14	d	6	l	4	bY		
25	m	13	g	5	t	3	pY		

The upper bound for the number of diphones required for a given language is given by the square of the size of the phoneme inventory (including the 'pause phoneme'). There are 38 phonemes in the corpus, including pause, so the upper bound for the number of intra-word and inter-word diphones required for a full corpus is 38 squared, i.e. 1444. A total of 253 diphones are represented in the 'MarketSpeak' corpus, i.e. about 20% of the upper bound. The diphones, sorted by frequency, are shown in Table 4. Even if the constraints on Igbo phoneme co-occurrence are such that the full complement of 1444 diphones is not attested, the number of unique phoneme occurrences shows that the entire potential of the Igbo diphone set is not reached. It is therefore clear from the frequency tables for the corpus that, while sufficient for microvoice testing purposes in the selected scenario, the current diphone coverage of this corpus will not provide a complete Igbo voice.

Table 4: Igbo diphone frequency list from 'MarketSpeak' corpus.

Freq	Diphone	Freq	Diphone	Freq	Diphonne	Freq	Diphone	Freq	Diphone
28	a _	3	o r	2	i Nw	1	pY U	1	h E
19	_ a	3	O m	2	I n	1	pY O	1	h a
16	U _	3	o l	2	I m	1	pY o	1	gw u
12	O _	3	O k	2	I k	1	p i	1	gw O
12	_ o	3	n E	2	i e	1	O z	1	gw e
12	e _	3	m U	2	I d	1	o v	1	gw a
11	_ O	3	l e	2	h I	1	O tS	1	G I
11	I _	3	J a	2	h i	1	O pY	1	g e
10	_ U	3	I tS	2	g I	1	o pY	1	e z
10	i _	3	g U	2	g i	1	O p	1	e w
10	E _	3	g o	2	g a	1	o p	1	e r
10	_ e	3	E r	2	_ g	1	o o	1	e O

Freq	Diphone		Freq	Diphone		Freq	Diphonne		Freq	Diphone		Freq	Diphone	
8	m	a	3	dZ	I	2	f	O	1	O	l	1	E	Nw
8	_	m	3	d	U	2	f	E	1	o	k	1	E	kw
8	k	a	3	b	I	2	e	n	1	O	h	1	E	k
8	a	n	3	a	J	2	E	m	1	O	f	1	e	k
7	m	m	2	U	z	2	e	h	1	n	z	1	e	i
7	_	i	2	U	t	2	e	g	1	NX	U	1	e	gw
6	z	U	2	U	r	2	E	f	1	NX	a	1	E	G
6	_	u	2	U	n	2	e	dZ	1	_	Nw	1	E	g
6	u	_	2	u	n	2	E	d	1	n	tS	1	e	d
6	_	n	2	U	bY	2	d	e	1	n	r	1	e	a
6	_	I	2	tS	a	2	a	r	1	n	I	1	dZ	u
6	_	E	2	t	e	2	a	p	1	n	gw	1	d	u
5	o	_	2	s	e	2	a	NX	1	m	e	1	d	o
5	kw	a	2	r	U	2	a	kw	1	m	b	1	d	i
5	k	I	2	r	a	2	a	I	1	m	_	1	d	E
5	I	a	2	o	m	2	a	h	1	l	u	1	d	a
5	dZ	i	2	O	kw	1	z	u	1	kw	U	1	bY	u
5	a	z	2	O	g	1	z	O	1	kw	E	1	bY	O
4	U	O	2	o	g	1	z	I	1	kw	e	1	bY	e
4	r	O	2	O	dZ	1	z	i	1	k	U	1	bY	a
4	O	r	2	O	d	1	z	a	1	k	O	1	b	U
4	n	U	2	Nw	E	1	w	u	1	J	E	1	b	O
4	i	k	2	Nw	a	1	v	a	1	j	a	1	b	E
4	I	b	2	n	u	1	u	s	1	_	J	1	b	e
4	_	dZ	2	n	n	1	u	r	1	I	z	1	a	t
4	d	I	2	n	kw	1	u	l	1	I	S	1	a	pY
4	a	tS	2	n	g	1	U	kw	1	i	s	1	a	O
4	a	k	2	n	d	1	U	I	1	I	r	1	a	m
3	U	k	2	n	a	1	u	h	1	i	r	1	a	l
3	u	d	2	m	j	1	U	g	1	I	O	1	a	j
3	t	U	2	m	I	1	U	d	1	I	kw	1	a	gw
3	tS	O	2	l	a	1	u	bY	1	i	kw	1	a	g
3	tS	I	2	kw	O	1	U	b	1	i	i	1	a	f
3	s	i	2	k	o	1	u	b	1	I	h	1	a	E
3	r	o	2	k	E	1	U	a	1	i	gw	1	a	dZ
3	r	i	2	k	e	1	tS	i	1	i	d	1	a	a
3	r	E	2	_	k	1	S	a	1	i	bY	1	_	_
3	p	a	2	j	O	1	r	u	1	i	a			
3	o	s	2	i	t	1	r	I	1	h	u			

In order to create a complete voice, a number of strategies are available for creating the required data set:

1. *Wordlist Strategy*: Create a list of words and word combinations which exhaust the number of possible combinations of 2 phonemes, and place these in a context frame corresponding to the traditional "Say ___ again." This traditional method has many disadvantages, for instance creating artefacts resulting from boredom and ordering effects.
2. *Natural Context Strategy*: Create a set of sentences containing at least one token of all possible combinations. This method has the advantage of more realistic and interesting prompts and is less likely to contain artefacts of the kind noted under the Wordlist Strategy.
3. *Minimal Natural Context Strategy*: From the set of sentences defined for the second strategy, select the smallest set containing all the diphones (Bachan 2010). It may be possible to reduce the number of prompts by this method to only a few hundred. Given a large corpus of sentences, this can be done with a software tool; final checking for missing diphones is

always necessary. This method has the advantage of producing a smaller corpus than the Natural Context Strategy for annotation, a time-consuming and expensive activity, even if automatic segmentation with post-editing is available.

The main approach taken here is the Wordlist Strategy, due to the small size of the scenario-determined vocabulary, supplemented with the Minimal Natural Context Strategy for multi-word expressions.

Duration specifications

The calculation of appropriate durations for sounds in different contexts is a complex issue, and was dealt in a highly simplified fashion in the present work by simply automatically averaging the durations for each phoneme as specified in the annotation time stamps, or, in the case of close copy tests, by automatically copying the durations from the annotation time stamps. A kind of null case was also generated, using arbitrary uniform segment lengths of around 80ms, which surprisingly led to very comprehensible and reasonably natural utterances. On reflection, this initially surprising result could, however, be a function of what is apparently a fairly uniform syllable timing in Igbo, another empirical question for future work.

Pitch specifications

For close-copy tests there was no problem in automatically copying the sampled F0 from the original recordings, averaging the samples, and using these in the PHO representations. The results were pretty much identical to the original recordings.

A null case was also generated, based on the frequency model introduced above, retaining the asymptotically descending terracing function:

$$F0(phoneme_i) = f(baseline, onsetpitch, declination^i)$$

This null case is 'tone-deaf', i.e. with no modification by lexical or grammatical High, Downstepped High and Low tone. Consequently, the contour was predictably not very natural, but still comprehensible, relying on native speaker ability to disambiguate in the case of tonal minimal pairs for which the tonal cues were missing. The practical value of this null contour in the present tutorial context was initially to provide a starting point for manually adding local modifications of High tone, Downstepped tone, Low tone and utterance-final lowering, for training and evaluation purposes.

The automatic assignment of lexical and grammatical and intonationally determined pitch values is a front end issue and outside the scope of the present account; for test purposes, the values were calculated straightforwardly with a Python script.

Diphone database construction

Given the recordings and the annotations, a number of software tools were used to construct the diphone database (voice).

In order to extract the diphones from the recordings, a suite of Python scripts developed by the third author was used: the first creates a table of diphones and calculates their time-stamps from information in the annotations; the second creates the diphone file set and the table of diphone information required by the Mbrolator. The format is shown in Table 5.

Table 5: Input format for Mbrolator software tool set.

Diphone filename	Diphone		Start	End	Mid
a-ee_UgoIgbo01_599.wav	a	E	800	2549	1588
a-ii_UgoIgbo01_537.wav	a	I	800	2332	1526
a-jj_UgoIgbo01_336.wav	a	J	800	3073	1813
a-nnX_UgoIgbo01_269.wav	a	NX	800	2800	1882
a-oo_UgoIgbo01_592.wav	a	O	800	3030	1856
a-SIL_UgoIgbo01_4.wav	a	_	800	3751	2151
a-a_UgoIgbo01_526.wav	a	A	800	2365	1510
a-dzz_UgoIgbo01_322.wav	a	dZ	800	3358	2222
a-f_UgoIgbo01_469.wav	a	F	800	4486	2527
a-g_UgoIgbo01_182.wav	a	G	800	2652	1696

The Mbrolator tools were provided under licence by the developers, and consist of a library of signal processing routines, and three user accessible tools: one for setting basic conversion parameters, one for processing the parameters, and one for creating the database on the basis of these parameters. The input to the Mbrolator tools consists of the set of diphones in separate files, an information file containing phoneme specifications, and a table containing a list of diphones with start, mid and end time-stamps from the diphone files. The output is a database in the form of a single file containing the normalised diphones and information from the information file.

Full details of for use of the software tools are given on the MBROLA website and in Bachan (2007, 2010), Bachan et al. 2006 , Gibbon et al. 2008).

Conclusion and prospects

A toolset and a workflow designed for speech technology education purposes were created with the aim of filling a gap in the available applied linguistic and speech technology literature, with the practical goal of simplifying the creation of basic synthetic voices for restricted scenario speech synthesis applications for under-resourced languages. On this basis, a working prototype DSP component for Igbo speech synthesis was created and satisfactorily evaluated using standard methods (Gibbon et al. 1997).

The choice of a traditional diphone synthesiser was motivated in detail, and the essential stages of the workflow were presented. For the Igbo ‘MarketSpeak’ scenario a set of prompts was created, recorded and annotated, and diphones and metadata about the diphones were extracted automatically from the annotated recordings, and a voice was created.

As noted during the discussion, the NLP component and its computational linguistic foundations were only treated in passing. There are three main descriptive linguistic and computational linguistic complexities which are not accounted for by the models discussed here: grammatical and lexical tone mapping, segment-tone interaction, and influence of sentence-level intonational categories such as focus (which can disrupt local downstep) and questions (which can disrupt overall declination); cf. Ikekeonwu (1993). There are models available which appear to be suitable for these purposes, such as multi-tape finite state transducers, which have been used for related languages (Gibbon et al. 2006b), and were sketched in Figure 3. These issues remain for future work.

Many attempts have been made to provide natural language and spoken language resource and toolkit specifications for under-resourced languages, as in the BLARK, the Basic LAnguage Resource Kit (Krauwer 2005), and extensions such as those of Gibbon et al. (2006a). However, these resource kits have not been specifically designed for use in training and basic development situations by linguists and computer scientists with no intensive speech engineering training, but with the plain motivation to create practical speech synthesis applications for their languages.

There have also been many specific state-of-the-art speech technology applications created for previously under-resourced languages in many countries across the world, including African languages. These cutting edge applications represent great strides forward in the field, and in specific application areas, but so far are only usable by specialists. There have been initiatives for combining cutting edge methodology with ease of use, such as the SPICE system (Schultz et al. 2007). However these are concept studies, and not generally available, and their dependence on using servers via the internet make them unsuitable for many places with slow internet connections or no internet connection at all. Consequently, there is a real need and a place for tutorial educational activities focussing on well-tried traditional methodologies such as those discussed in the present report. Bearing in mind the numbers of less resourced languages which do not yet have advanced technological infrastructure, there will continue to be a place for such tutorial educational methods for a long time to come.

References

Akinlabi, A. & M. Liberman. 2000. "Tonal Complexes and Tonal Alignment". *North East Linguistics Society* (*NELS*) 31.

Bachan, J. 2007. Automatic Close Copy Speech Synthesis. *Speech and Language Technology.*

Volume 9/10. edited by Grażyna Demenko. Poznań: Polish Phonetic Association. 107-121.
Bachan, J. 2010. Efficient diphone database creation for MBROLA, a multilingual speech synthesiser. *XII International PhD Workshop (OWD 2010). Conference Archives PTETiS,* 28. 23-26 October 2010. 303-308.
Available: http://mechatronika.polsl.pl/owd/pdf2010/JBachan.pdf
Accessed: 2011/03/29.
Bachan, J. & D. Gibbon. 2006. Close Copy Speech Synthesis for Speech Perception Testing. *Investigationes Linguisticae, vol. 13*, pp. 9-24.
Available:
http://www.staff.amu.edu.pl/~inveling/pdf/Jolanta_Bachan_Dafydd_Gibbon_INVE13.pdf
Accessed: 2011/03/29.
Bloomfield, L. 1933. *Language*. New York: Henry Holt.
Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5:9/10, 341-345.
Boersma, P. & D. Weenink. 2011. Praat: doing phonetics by computer [Computer program]. Version 5.2.21.
Available: http://www.praat.org/
Accessed on: retrieved 29 March 2011.
Connell, B. 2002. Downdrift, downstep, and declination. *Proc. TAPS (Typology of African Prosodic Systems Workshop), Bielefeld, Germany, March 18-20 2001*. edited by U. Gut & D. Gibbon. Bielefeld: Universität Bielefeld.
Available: http://www.spectrum.uni-bielefeld.de/TAPS.
Accessed: 2011/03/29.
Duruibe, U. V. 2010. *A Preliminary Igbo text-to-speech application*. BA thesis. Ibadan: University of Ibadan.
Dutoit, T. & V. Pagel. 1996. Le projet MBROLA : vers un ensemble de synthetiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. *Actes des Journees d'Etudes sur la parole, Avignon*. 441-444.
Dutoit, T. 1997. *An Introduction to Text-To-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.
Gibbon, D. 1987. Finite State Processing of Tone languages. *Proc. EACL '87, Copenhagen*. 291~297.
Gibbon, D. 2001. Finite state prosodic analysis of African corpus resources. *Proc. Eurospeech 2001, Aalborg, Denmark*, Vol. I:83-86.
Gibbon, D. 2009. Prosodic Rank Theory: on the formalisation of prosodic events. In P. Łobacz, P. Nowak & W. Zabrocki, eds. *Language, Science and Culture. Essays in Honor of Professor Jerzy Bańczerowski on the Occation of his 70th Birthday*. Poznań: Adam Mickiewicz University Scientific Press. 93-126.
Gibbon, D., R. Moore & R. Winski. (Eds.) 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.
Gibbon, D., I. Mertins & R. Moore. (Eds.) 2000a. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers.
Gibbon, D., E.-A. Urua & U. Gut. 2000b. How low is floating low tone in Ibibio? *Proc. 30th Conference on African Languages & Linguistics, U Leiden, August 2000*. 27-30.
Gibbon, D., F. Romani Fernandes & T. Trippel. 2006a. A BLARK extension for temporal annotation mining. *Proc. LREC 2006, Genoa*.
Gibbon, D., E.-A. Urua & M. Ekpenyong. 2006b. Problems and solutions in African tone language Text-To-Speech. Justus Roux, ed., *Proc. Multiling 2006 Conference, Stellenbosch, South Africa*.
Gibbon, D. & J. Bachan. 2008. An automatic close copy speech synthesis tool for large-scale speech corpus evaluation. K. Choukri, ed., *Proc. LREC 2008, Marrakech, Morocco*. Paris: ELDA. 902-907.

Gibbon, D., P. K. S. Pandey, D. M. K. Haokip & J. Bachan. 2009. Prosodic issues in synthesising Thadou, a Tibeto-Burman tone language. *Proc. Interspeech 2009, 6-10 September 2009, Brighton, UK*.

Hirst, D. & A. di Cristo. (Eds.) 1998. *Intonation Systems. A Survey of Twenty Languages*. Cambridge: Cambridge University Press

Ikekeonwu, C. I. 1993. Intonation and Focus: A Reanalysis of Downdrift and Downstep in Igbo. *Lund University, Dept. of Linguistics Working Papers* 40, 95-113.

Krauwer, S. 2005. ELSNET and ELRA: a common past and a common future. *ELRA Newsletter*, 3:2.
Available: http://www.elda.org/article48.html (14.10.2005 06:27:33) .
Accessed: 2011/03/29.

Liberman, M, Y. & J. B. Pierrehumbert. 1984. Intonational Invariance under Changes in Pitch Range and Length. *Language Sound Structure*, edited by M. Aronoff and R.T. Oehrle. MIT Press. 157-233.

Roux, J., P. Scholtz, D. Klop, C. Povlsen, B. Jongejan & A. Magnusdottir. 2010. Incorporating Speech Synthesis in the Development of a Mobile Platform for e-learning. *Proc.LREC 2010, Valetta, Malta*.

Schultz, T, A. W Black, S. Badaskar, M. Hornyak & J. Kominek. 2007. SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems. *Proc. Interspeech 2007, Antwerp, Belgium, August 2007*.