

A Hybrid Text-To-Speech system for Afrikaans

Francois Rousseau and Daniel Mashao

Department of Electrical Engineering, University of Cape Town, Rondebosch,
Cape Town, South Africa, frousseau@crg.ee.uct.ac.za, daniel.moshao@ebe.uct.ac.za

Abstract – A high quality text-to-speech (TTS) system must have the following attributes: flexibility, naturalness, pleasantness and understandability. There are two popular techniques that are used to design TTS systems: unit selection synthesis and diphone concatenative synthesis. Limited domain unit selection synthesis is a unit selection technique with a restricted vocabulary. This technique produces very natural, pleasant and understandable synthetic speech but lacks in flexibility. The diphone concatenative synthesis technique on the other hand produces very flexible speech synthesis but lacks in naturalness, pleasantness and understandability. In this paper we design a hybrid TTS system that combines these two techniques. We evaluate how the combination performs versus the requirements of an ideal TTS system. Results show that the pleasantness and naturalness of the system is above satisfactory and that the synthetic speech is easily understandable. Results can be improved by using a professional speaker.

Index terms: unit selection synthesis, diphone concatenative synthesis, text-to-speech

I. INTRODUCTION

THE quality of a text-to-speech system depends on its flexibility, naturalness, pleasantness and understandability [1]. Flexibility is the potential of the system to synthesize any possible word in the language [2]. Naturalness is how close to real speech the output of the system is [1]. Pleasantness is how pleasant the voice is [3] and understandability is how easy it is to understand the message when listening to it for the first time [1]. Even though there are a number of systems available in practice most of them can not satisfy all these needs. One popular system is the Festival Speech Synthesis System [4]. The system was designed in the Centre for Speech Technology Research (CSTR), at the University of Edinburgh, Scotland. It is an open source system with the ability to be a workbench for the development of new text-to-speech systems [5]. Festival is based on concatenative speech synthesis which is a technique that connects prerecorded units of speech derived from natural speech for synthesis [2, 6, 7].

Two types of concatenative speech synthesis are available with Festival. The first, called diphone concatenative synthesis

(DCS) produces very flexible speech synthesis, but lacks in naturalness, pleasantness and understandability [6, 7]. Diphones are simply all possible phone-to-phone transitions for a particular language with the square of the number of phones being the number of diphones present in that language [2]. Diphones give a very good coverage of all possible sounds in the given language and are the easiest units to join for speech synthesis [2, 7].

The second form of concatenative synthesis is called unit selection synthesis (USS). Limited domain (ldom) unit selection synthesis is the USS technique used for this work. It produces very natural, pleasant and understandable synthetic speech but it lacks flexibility since it can only synthesize words in a given vocabulary or database [8, 9].

The goal of this work is to combine the advantages of these two techniques into a high quality TTS system for the South African language, Afrikaans. The proposed system is illustrated in Figure 1.1 below.

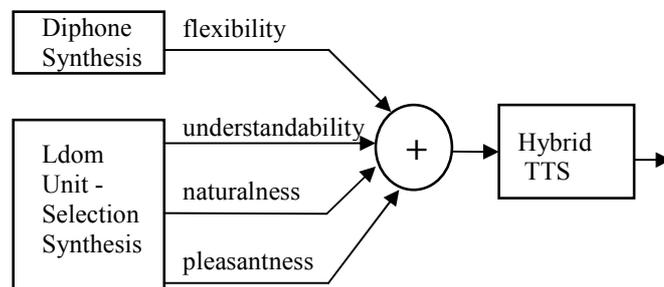


Figure 1.1: Hybrid TTS System

Afrikaans is the home language to approximately six million people in South Africa [10]. A previous TTS system for the language was built by SUN (University of Stellenbosch). The system is called the AST (African Speech Technology) project which is used in hotel reservation booking system [11].

Section 2 of this paper discusses the construction of the DCS system, the USS system and the combination of the two. Section 3 discusses the procedure used for testing the system. Section 4 shows experimental results while Section 5 gives conclusions based on the results.

II. CONSTRUCTION OF THE HYBRID TTS SYSTEM FOR AFRIKAANS

Like all TTS systems the hybrid system must have a front-end and a back-end. The front-end will be used for high-level synthesis while the back-end will be used for low level synthesis (see [2] for more detail). This is better illustrated in Figure 2.1

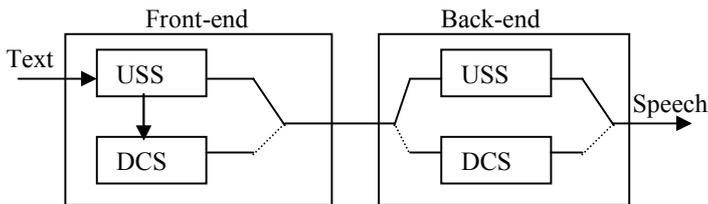


Figure 2.1: Front-end and Back-end of hybrid TTS system

Within the front-end the system decides whether to use the USS system or the DCS system for synthesis. The condition for using the DCS system is that if an unknown word (out of vocabulary word) is picked up within the front-end of the USS system then the system will switch to the DCS system [5, 8]. Therefore the DCS system is used as the back-up voice to the USS system.

The objectives for the proposed system are as follows

1. The system must be in the Unit Selection Synthesis (USS) system at all times.
2. We want the system to show its flexibility by falling back onto the DCS system when a particular word is unknown to the USS system's database.
3. We want the system to revert back to the USS voice after synthesis was carried out by either voice.

Figure 2.2 shows a flow diagram of these objectives.

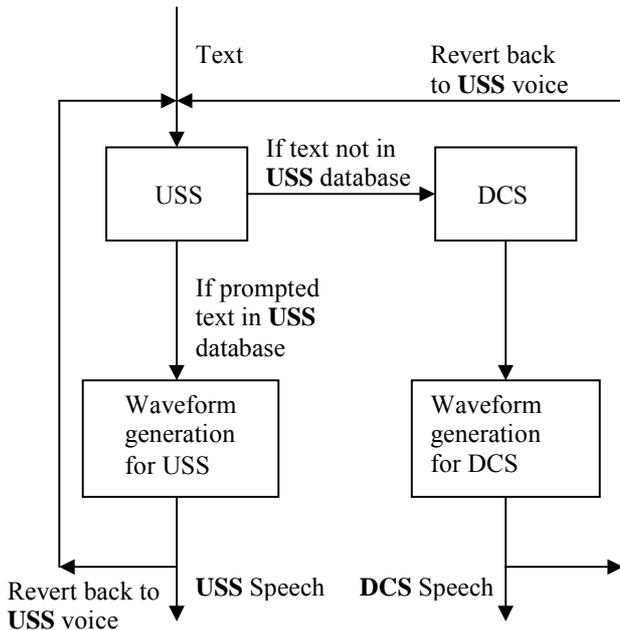


Figure 2.2 Flow diagram showing objectives of work

Once the objectives for the final system were clear the Festival Speech Synthesis System was employed to build both the proposed system [4].

2.1 Building the DCS system

This section describes the steps involved in building the diphone concatenative synthesis system.

Constructing the diphone database

The diphone database for Afrikaans was constructed using *Die Groot Woorde Boek*, Afrikaans dictionary [12]. In total 64 phones were found therefore 4096 diphones are present in the system. The database was generated by the system using the phone-to-phone (consonant-consonant, consonant-vowel, vowel-vowel and vowel-consonant) transition rules for Afrikaans. These diphones are then placed within non-sense words which are to be used for the extraction of the speech units for concatenation. Table 1 shows a list of diphones located within non-sense words.

Table 1: Examples of diphones located within non-sense words

| Diphones | Non-sense word | Diphones with-in non-sense word |
|---------------|----------------|---------------------------------|
| 'b-a' 'a-b' | Tababa | t a-b-a-b a |
| 'sj-a' 'a-sj' | Takasjata | t a k a-sj-a t a |
| 'kn-o' 'o-kn' | Takoknota | t a k o-kn-o t a |
| 'tj-e' 'e-tj' | Taketjeta | t a k e-tj-e t a |

Recording the speaker

This step allows for a uniform set of diphone pronunciations for the database. Recording was done using *na_record*, part of Edinburgh *speech_tools-1.3* package [15]. This recording system creates wave files of the recorded non-sense words and places them into a log file storing as *.wav files.

Labeling the non-sense words

The labeling of non-sense words identifies the positions of diphones within non-sense words. At minimum the start of the preceding phone to the first phone in the diphone, the changeover and the end of the second phone should be labeled [5]. Festival provides an automatic labeler called *make_labs* to automatically label the diphones. The results of this labeling algorithm are unsatisfactory since many of the diphones are mislabeled. Therefore the entire labeling of the diphone database had to be hand corrected, which was a very tedious task.

Building the diphone index

The diphone index is needed for the extraction of diphones from the acoustic non-sense words. The index is built by taking the diphone list and finding the occurrence of each diphone in a label [5]. For synthesis only the transition from the first phone to the second phone is required. Therefore the diphone will be extracted from the middle of the first phone to the middle of the second phone for synthesis.

Extracting parameters for synthesis

The Festival speech synthesis system uses a technique called Residual Excited Linear-Predictive Coding (residual LPC) for the re-synthesis of diphones as its synthesis technique. The technique is based on the assumption that a current speech sample $x(n)$ can be predicted from a finite number of previous p amount of samples $x(n-1)$ to $x(n-k)$ by a linear combination with an error $e(n)$ [6]. This error term is the residual signal. Therefore

$$x(n) = e(n) + \sum_{k=1}^p a(k)x(n-k), \quad (1)$$

and

$$e(n) = x(n) - \sum_{k=1}^p a(k)x(n-k) = x(n) - \tilde{x}(n) \quad (2)$$

where $\tilde{x}(n)$ is the predicted value, p is the linear predictor order and $a(k)$ are the linear prediction coefficients which are found by minimizing the sum of the squared errors over a speech frame [6]. This step therefore entails the extraction of the LPC coefficients and the LPC residuals from each diphone for re-synthesis. The following steps were carried out to perform this task.

(i) Extracting the pitchmarks

Residual LPC is a pitch synchronous technique that requires information regarding the positions of the pitch periods in an acoustic signal for synthesis. For this reason the pitchmarks in each speech waveform must be extracted [5]. Ideally the use of a laryngograph (measures the electrical activity in the glottas) would produce very accurate positions of the pitchmarks. For this research a laryngograph was not available and hence pitchmarks were extracted from the raw waveforms of each diphone.

(ii) Power normalization

It is difficult to achieve or access an ideal recording environment where there is no background noise, no natural environmental changes and no human fatigue. Therefore using a laboratory for recording introduces the factor of power fluctuations in the recorded non-sense words. This plays a big role in producing bad synthesis [5]. To overcome this power normalization was done on all the recorded non-sense words. The method used finds the mean power for each vowel in each of the non-sense words and then finds the power factor with respect to the overall mean vowel power [5].

(iii) Building the LPC parameters

Using the normalized power factors and the extracted pitchmarks, the LPC coefficients and residuals for LPC analysis were generated. The LPC coefficients were obtained using the speech tools program *sig2feat* (signal to feature vector). The LPC residuals were obtained using the speech tools program *sigfilter* which finds the residuals by inverse filtering the non-sense words [7].

Building lexicon support database

The lexicon support database consists of the letter-to-sound rules and pronunciation guides for the DCS system. Unpronounceable words and abbreviations are defined here. Some phones and diphones are not always as required when trying to pronounce certain words. Take the word “*Francois*” as an example. The first syllable of the word can be pronounced just by using the information of the phones. The second syllable is not pronounced correctly in the context of how the full word should be pronounced. For this reason the system needs to be told how to pronounce this syllable. Below is an example taken from the lexicon database that show how the syllable is pronounced [2].

```
(lex.add.entry
  ('Francois' nil (((f r a n) 0) ((s w a) 0))))
```

Now the system has a definition of how the word “*Francois*” should be pronounced and will be used at synthesis.

2.2 Building the USS system

Described here are the steps involved in designing a limited domain (ldom) unit selection synthesis system.

Setting up the back-up/prompt voice

As mentioned before the DCS system is used as the back-up voice to the USS system. Within the skeleton modules of the USS system there is an option of setting a *closest_voice* function that will call be upon as a back-up voice to the USS system in case it fails [5, 8]. The *closest_voice* for this work was set to be the DCS system. Now out of vocabulary words can be attempted to be synthesized showing the flexibility of the entire system. Not only is the *closest_voice* used for the task mentioned above, but also for the task of setting up the basic recording and labeling prompts for the USS system [5].

Designing the prompts

This step involves the defining the vocabulary of the USS system by filling it with sentences in the form of text. These sentences are placed within a prompt file which is to be used by the sub-processes used in building the system. Some examples are shown below

```
(time0001 “nul, een, twee, drie, vier, vyf, ses, sewe, agt, nege”)
(time0002 “goeie, more, dames, en, minere”)
(time0003 “jou, telefoon, nommer, is”)
(time0004 “jou, identiteits, nommer, is”)
```

The terms *time000** in front of each sentence is used to label the sentence so that it can be identified by the sub-processes [5]. In total twenty five sentences were defined which gives a limited but good coverage of a range of words in the Afrikaans language.

Recording the prompts

The same steps and tools used to record the diphones for the DCS system were used to record the vocabulary of the USS system. The only difference here was that we recorded the full words instead of units of words (diphones). The reason for the commas between consecutive words was to satisfy the recording strategy used in recording the vocabulary. This strategy ensures that there will be no overlapping of phones from consecutive words. This means that when a word is called for synthesis only that particular word will be synthesized and nothing else.

Labeling the vocabulary

As in the case of the DCS system the labeling of the recordings identifies the speech units used for synthesis. In this case the labels identify the diphones within the words of the sentences. These diphones are found between the commas labeled as pauses between the words within the sentences. The automatic labeling algorithm *make labs* again produced the problem of mislabeling and therefore hand corrections of the labeling errors had to be done.

Extracting parameters for synthesis

The same reason why the synthesis parameters are needed for the DCS is the reason why it was needed for the USS system. The same pitchmark extraction technique used for the DCS system was used for the USS system. Since both systems were recorded in the same environment power normalization had to be done on the USS system as well. The Festival module *simple_povernormalize* was used to normalize the power levels of the speech data for the USS system [5]. Then the pitch synchronous MELCEP parameters of the speech had to be generated using *make_mcep* provided by Festival [5].

Building the cluster units for the USS system

The next step was to build clusters of each unit in the database that appears more than once [5]. A target cost is then used to determine the correct or appropriate unit for synthesis [5, 8]. For this reason the USS system is also referred to as a clunit (cluster unit) synthesizer [5, 8].

III. TESTING PROCEDURE

This section describes the procedure that was used for testing the Hybrid TTS system. It discusses the Mean Opinion Score (MOS) rating system that was used to score the system, why it was chosen to use such a system and the evaluation sheet designed to get the opinions of the test subjects.

MOS rating systems have been proven to be a reliable evaluation technique for opinion tests [3]. It was decided to follow [3] in using a 6 point scoring system instead of the usual 5 point system [13]. This prevents neutral scores in the middle of the point scale.

Figure 3.1 shows the evaluation sheet use to evaluate the system. It is based on the ITU MOS questionnaire [14]. The following questions were asked:

1. How pleasant is the voice you just listened to?
2. How much listening effort is required to understand what was said?
3. How natural is the voice you just listened to?
4. Overall impression of the system

| Hybrid Text-To-Speech System for Afrikaans Evaluation Sheet | |
|--|---|
| Date: | |
| Home Language: | |
| 1. How Pleasant is the voice you just listened to? 6. Very pleasant 5. Pleasant 4. Satisfactory 3. Tolerable 2. Unpleasant 1. Very unpleasant | 2. How much listening effort was required to understand what was said? 6. No effort required at all 5. Minimum effort required 4. Fair 3. Decent amount of effort 2. Maximum effort required 1. Can not understand the |
| 1. How natural is the voice you just listened to? 6. Very natural 5. Natural 4. Satisfactory 3. Tolerable 2. Unnatural 1. Very unnatural | 1. Overall impression of the system 6. Excellent 5. Good 4. Fair 3. Tolerable 2. Poor 1. Horrible |

Figure 3.1: Hybrid TTS system evaluation sheet

Ten subjects (five Afrikaans and five English) were used to listen to five different sentences. The test sentences were:

1. Nul een twee drie vier vyf ses sewe agt nege tien.
2. Goeie more dames and minere.
3. Welkom by die demonstrasie van 'n Afrikaans rekenaar stelsel.
4. Dit is die einde van hierdie demonstrasie.
5. Dankie dat U geluister het, totsiens.

Each sentence was played once only. After all five were played the subjects were asked to give there opinions according the evaluation sheet above. The scores for each question were then averaged and results are shown and discussed in the next section.

The flexibility of a TTS system can not be tested using subjective listening tests, and therefore this system only shows its flexibility by invoking the DCS system when ever a word is not in the vocabulary of the Ldom USS system.

IV. EXPERIMENTAL RESULTS

This section shows the results of the testing procedure used for evaluating the proposed system. The results for each question were averaged and tabulated and is shown in Table 2.

Table 2: Mean Opinion Score ratings for the system

| Question | Mean Opinion | Meaning of score |
|--------------------|--------------|--------------------|
| Pleasantness | 4.3 | Above satisfactory |
| Understandability | 5.6 | No effort required |
| Naturalness | 4.3 | Above satisfactory |
| Overall Impression | 4.7 | Good system |

According to table 2 the system has

1. A mean pleasantness of 4.3. This means that the mean opinion of the subjects were that the system is between pleasant and satisfactory. Compared to [3] this is a good results since it is above average. This result is dependant on the actual voice used for building the system. Using professional speakers will improve this result since their voices are generally of a higher quality than the average speaker.
2. A mean understandability of 5.6. Probably the most important result is the understandability of the synthetic speech. This conveys whether the correct message conveyed or not.
3. A mean naturalness of 4.3. This proves that pleasantness and natural are both dependants on the voice talent used to build TTS systems. Once again this result can be improved by using a professional speaker.
4. A mean overall impression of 4.7. According to the MOS rating system this indicates that it is a good system and is above average [3].

V. CONCLUSIONS

In conclusion it can be said that with an above satisfactory pleasantness and naturalness, a high understandability and a good overall impression that the hybrid TTS system for Afrikaans is a good system and that it could be used in practical systems. This is supported by the high rating of the understandability of the system and the general opinions of the subjects. The pleasantness and naturalness are opinions on the voice talent used to build the system and can be improved by using a professional speaker.

REFERENCES

- [1] "Assessing Text-to-Speech System Quality", White Paper, SpeechWorks International, Available at http://www.tmaa.com/tts/white_papers.htm
- [2] F. Rousseau, Dr. D.J Mashao, "Increased Diphone Recognition for an Afrikaans TTS system", Proceedings of PRASA 2004, pp 113-117, Cape Town
- [3] G. P. Sonntag, T. Portele, F. Haas and J. Kohler, "Comparative Evaluation of Six German TTS Systems", Proceedings of Eurospeech 1999, Vol. 1, pp 251-254, Budapest
- [4] A. W. Black, R. Clark, K. Richmond, S. King "The Festival Speech Synthesis System", University of Edinburgh, Scotland www.csrt.ed.ac.uk/projects/festival, Last accessed 15 October 2004
- [5] A. W. Black, K. Lenzo "Building Synthetic Voices", unpublished document, Carnegie Mellon Universtiy, Available at <http://festovx.org.bsv>
- [6] S. Lammety, "Review of Speech Synthesis Technology", Master's Thesis, Department of Electrical Engineering, Helsinki University of Technology, March 1999, Available at <http://www.acoustics.hut.fi/~slemment/dipp/index.html>, Last accessed 5 August 2004
- [7] N. Rochford, "Developing a new voice for Hiberno-English in The Festival Speech Synthesis System", Final Year Thesis Project, Trinity College Dublin. Available at <http://www.cs.tcd.ie/courses/csll/projects4.html>, Last accessed 7 June 2004
- [8] A. Schweitzer, N. Braunschweiler, T. Klankert, B. Möbius, B. Säuberlich, "Restricted Unlimited Domain Synthesis", Proceedings of Eurospeech 2003, pp 1321-1324, Geneva
- [9] B. Langner, A. W. Black, "Creating a Database of Speech In Noise For Unit Selection Synthesis", 5th ISCA Speech Synthesis Workshop, PiTTsburgh, USA 2004
- [10] CENSUS 2001, "Statistics South Africa", Online resource: <http://www.statssa.gov.za/census01/html/default.asp>, Last accessed April 2005
- [11] Prof J. Roux, Prof L. Botha, Prof J du Preez "African Speech Technology", Online Resource: www.ast.sun.ac.za, Last accessed 7 October 2004
- [12] Kritzenbeurg, M. S. B.(Matthys Stefanus Benjamin), "Groot Woordeboek", Pretoria, Vanschaik 1972
- [13] ITU-T Recommendations P.85 "A method for subjective performance assessment of the quality of speech output devices", International Telecommunications Union publication, 1994
- [14] Mahesh Viswanathan, Madhubalan Viswanathan, "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", Computer Speech & Language, Vol. 19, Issue 1, pp 55-83, 1 January 2005
- [15] P. Taylor, R. Caley, A. W. Black, S. King, "Edinburgh speech tools library system documentation", http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/, June 1999.

F. Rousseau completed his BSc (Elec. Eng.) in 2003. He is currently pursuing an MSc in Electrical Engineering at the University of Cape Town. This is his second year on study.

Dr. D. Mashao is a senior lecturer at the University of Cape Town and head of the Speech Technology and Research Group. He is also the supervisor of the above-mentioned author.

