M. Hood Supervised by A. Lobb and S. Bangay
G01H0708

Creating voices for the Festival speech synthesis system.

Abstract

This project focuses primarily on the process of creating a voice for a concatenative Text-To-Speech (TTS) system, or altering the TTS systems own standard output voice to sound more like the target voice. The main system used is the Festival Speech Synthesis System developed by the Center for Speech Technology Research at Edinburgh University.

Festival is an open source research project and was selected as it offers the use of multiple languages and multiple voices in those languages. The process of voice creation requires a speaker to record many diphone sounds, which can take a few hours. This is then followed by the processes of labelling, pitch mark extraction and building, which while there are automation tools to help, is still a very labour intensive process and very time consuming. The voice will, even after this, still generally be of a fairly low quality, and much more work is required to get an exact copy of a speaker's voice. The project looks at first understanding this process and then at ways in which it can be sped up or automated.

There is also the process of voice adaptation or transformation to consider. This is where, instead of recording a new voice from scratch, the speaker reads a relatively short segment of text and this is compared to an already existing voice and their differences noted. This allows the creation of a filter that can be put on the original voice to make it sound more like the target voice. There are a number of different ways of achieving this, but the source + filter approach is the most common in recent years. This process has a number of advantages most notably the dramatic reduction of time required to produce a new voice, and at fairly comparable output quality, especially if the built voice does not need to be exactly like the speaker. However most work has been done in adapting human voices. There has been very little work in transforming a TTS's output voice. Voice adaptation also has the bonus of taking

up much less space as there is no need to store the new diphone sound files which can be as large as a few hundred megabytes.

Literature Review

1. Introduction

This literature review looks at voice creation for a speech synthesis system, specifically the Festival speech synthesis system. The paper discusses speech synthesis, looking at its past and some of the current research being done in the field. The main focus however is on the process of creating a voice from scratch, for Festival, and then look at voice adaptation as a possible alternative. The final goal of the project is to allow the easy creation of many voices for use in an E-Book reader; the final section covers a similar story teller.

2. History of Speech Synthesis

Speech synthesis has been around for a number of years. It really became a serious research project in the mid 80s. A number of academic and commercial projects have been underway for many years. A lot of the early research done in the area of text-to-speech synthesis is explained and demonstrated in *Review of text-to-speech conversion for English* [Klatt, 1987]. The paper also comes with recordings of a number of the systems to better show how they work. The paper outlines how long the different types of TTS systems have been around and what the advantages of each approach are. These early approaches vary quite considerably, as various methods have been attempted over the years. Currently the most popular way of performing text-to-speech translation is to use a concatenative system, where individual sounds or diphones are strung together to form the speaking.

3. Current Speech Synthesis Systems

There are many uses for TTS system, these include: digital actors, avatars, digitised speech (in conjunction with speech recognition and possibly a translator), next generation chat applications. As such there is a lot of research in the area. The following are some of the better TTS systems.

Commercial:

- *Whistler (**W**indows **H**ighly **I**ntelligent **ST**ochastic ta**L**k**ER**)* was developed by Microsoft in 1998 and today is used in a number of their applications including Narrator as part of Windows 2000 and XP. The over all quality of

the speech is not great and sounds like the typical talking computer. The project originally proposed a novel voice training system. Unfortunately this research appears to have been discontinued and no information about the training was released. The present system has no facilities to make new voices.

- *SoftVoice* developed by SoftVoice, Inc was first written in 1979 and is now released as a set of Windows DLLs to enable programmers too easily add speech synthesis capabilities to their products. SoftVoice gives decent voice quality, but lacks any programs or documentation on how to create a new voice.
- *Lucent TTS engine* was developed by Bell Laboratories. The engine is to be used in a number of Bell's own telecommunication and messaging products as well as being available to other developers. The system works primarily within the context of a telephone system and is obviously written with that in mind.
- *Natural Voices* is an ongoing research project at AT&T. The aim of the project is to get as natural as sounding speech for use in telephone systems. The quality of the output produced by the system is very impressive and has been used in a number of movies and recognised as a leading example by Discover Magazine in their article *The Mathematics of . . . Artificial Speech* [Discover Magazine, 2003]. At present there is no facility or documentation on new voice creation.

Academic:

- *Center for Speech Technology Research (CSTR)* at the University of Endinburgh is one of the leading research groups in the field of Text-To-Speech. It released the Festival speech synthesis system, the TTS used in this project. The CSTR also have the FestVox project, which is a project looking at voice creation for Festival. Festival uses the concatenative approach as do almost all of these systems. Its speech quality is good, but can be inexpressive. It is also not as clear as the Natural Voice project. Festival is designed to be easy to use as well as edit. It is fully open source and modules can easily be added to it. As such many other researchers use festival in their

own work. It also is one of the only products to actually look into creating voices.

- *Center for Spoken Language Understanding (CSLU)* is a research group at OGI School of Science and Engineering. This research group looks at a number of areas related to speech and have a detailed section covering speech synthesis. The group use Festival as one of their major components, unfortunately their section on voice creation is not available at this time.
- *Speech Research Lab* at the University of Delaware is a group interested in speech synthesis for use with disabled or ill patients, enabling them to better communicate with others. The group have developed the ModelTalker TTS system, the system produces output comparable to Festival, but is aimed more at the end user, it does not have the same level of control and customisation. There is also the InvTool program which is designed to help create new voices for ModelTalker. The system works the user repeating words spoken by ModelTalker and the diphones extracted from them. This makes the process much easier for a novice user and also by analysing each sound when it is recorded, before continuing onto the next one it means a lot of the extract effort put into cleaning up the sounds is done away with. The InvTool could be used as a good model to build a voice creation tool for Festival.

4. Festival and FestVox

For this project the speech synthesis system chosen is Festival. Festival is "a free, portable, language independent, run-time speech synthesis engine for various platforms under various APIs." [Black et al, 2003] As Festival is the main program used in this project one of the most important pieces of literature involved in the early stages of the project is *Building Synthetic Voices* [Black et al, 2003]. This work is part of the FestVox project which is involved in the creation of voices for the Festival system. The book is written by two of the creators of Festival but unfortunately is not finished at this time, and hence large sections are missing, the book however is invaluable to my work, even if just in the first stages.

Building Synthetic Voices [Black et al, 2003] gives an overview of speech synthesis systems. It explains the general anatomy of a synthesizer and how the engine is broken into three main sections namely

Text analysis

Converting the raw text into words and utterances.

Linguistic analysis

This is where the text is analysed for pronunciation and the words are assigned prosodic structure namely phrasing, intonation and duration.

Waveform generation

This is where the pronunciations are turned into the output waveform by extracting sounds from a database and then applying various filters.

A paper that was very helpful in this area is *A Short Introduction to Text-to-Speech Synthesis* [Dutoit, 1996]. This paper gives a much more detailed look at how a TTS system works, breaking the process down into Natural Language Processing (NTP) and Digital Signal Processing (DSP). NTP essentially involves the Text analysis and Linguistic analysis stages outlined above. The paper continues by discussing the DSP side, explaining the difference between rule based synthesizers and concatenative synthesizers. The difference being mainly that concatenative synthesizers possess a very limited knowledge of the data they handle. Words and utterances are generated from much smaller components. The paper also briefly deals with database preparation and how the individual sounds are stored. And then concludes that while there is still a long way to go present synthesizers should be capable of human like speech in the near future.

Building Synthetic Voices [Black et al, 2003] explains how to use the Festival system and then goes into detail about the processes involved in the creation of new languages and voices. It details the steps of recording, labelling, pitch mark extraction, building diphone list and linking it all together. The actual processes involve the use of the FestVox tools but even so still require a lot of effort on the behalf of the user. In my project one possible approach is to design a tool similar to

InvTool to help automate or at least speed up the process of creating a voice form scratch.

5. Voice Adaptation

Another approach to making new voices is the process of voice adaptation (also known as voice transformation). The basics of this approach are outlined in *Voice Transformation* [Tang, 2002]. The paper deals with voice in general and describes a number of properties that they "consider sufficient to characterise speech" [Tang, 2002], by emulating these properties it is possible to create a new voice. The Paper gives little more than an outline of what is involved in transforming a voice and no detailed process is given.

In the masters thesis *New methods for voice conversion* [Turk, 2000] the process of voice conversion is covered in great detail, along with an overview of an application developed as part of the project. While this thesis does not deal with voice adaptation for a TTS system, the processes outlined are fairly similar. The general idea is that two recordings are made, one of the source voice and the other of the target voice. By using mapping functions it is possible to work out how one voice differs from the other. From this it is possible to create a filter that will convert the source voice into the target voice. This means that any future source recordings can be made to sound like the destination voice. I propose that a similar process can be applied to the output of Festival when it is speaking in its standard voice, and hence although Festival is actually speaking in its normal voice, the filter will make the speech sound more like the target voice.

There are a number of ways of achieving this transformation, some of these require offline processing, but this is not really an option for use with a TTS system as the entire process should be real time. One approach often used when converting voices

is that of pitch shifting. Pitch shifting is the changing of the pitch of a waveform in real-time. It is used in many applications that deal with sound in general.

There are however a number of limitations to merely shifting the pitch, as it often does not leave the voice in a believable state. In an article released by TC-Helicon a Vocal Technologies group: *Pitch shifting and voice transformation techniques* [BASTIEN]. A few processes related to pitch shifting are discussed. The article discusses how we control our sound source (our vocal cords) and our resonator (our vocal tract) independently; hence merely shifting the pitch will not make the new voice sound realistic. The article briefly discusses what it calls more natural approaches. The voice is altered in two separate ways, by changing the source of the voice and the resonator that gives it its unique sound.

The sound of our speech is produced by a source (in our case our vocal cords) but it is the way in which each individual changes his resonator (vocal tract) that gives us our unique voices. This source + filter model is interesting and in many ways relevant to the process of changing a voice for a TTS system. If we consider Festival's standard output voice as the source and use a filter (resonator) to change the sound then it is possible to make festival talk in many voices, but only need to have one sound database.

An academic paper that also looks at the voice transformation problem from a more natural aspect is *A new voice transformation method based on both linear and nonlinear prediction analysis* [Seung Lee et al, 1995]. This paper uses the codebook approach; this is where there is a set of words to be read by the target voice and compared to the same words read by the source. The paper details how their process breaks the acoustic features into linear and non-linear parts and deals with these separately. The process is very detailed and the paper deals with the mathematics behind the problem. There is also a section at the end where the converted voices are

tested. The paper concludes "A listening test shows that the proposed method makes it possible to convert speaker's individuality while maintaining high quality" [Seung Lee et al, 1995]. This approach is not that suited to dealing with the problem of creating voices in Festival as Festival does not generally use a code-book. But the idea of using a listening test is useful in comparing the results produced by this project to other vice creation efforts, such as FestVox and InvTool.

The master's research project *Transforming Voice Quality and Intonation* [Gillet, 2003] looks at the creating new voices while taking into consideration the source + filter approach. It expands on work carried out by Kain et al, and discusses in detail how the quality and intonation aspects of voice conversion are dealt with separately. While this project is not specific to a TTS system, it was produced by the same department as the one that releases Festival. It is obvious that the method detailed in the project is applicable to the problem of creating a filter to a TTS system. With its intonation work, this project gives, if anything, too detailed an approach, this is due to the often monotone output of a TTS system. With the project there are also sound samples giving source, target and final voices. The results are promising, and my project would aim to replicate them for Festival's output as opposed to changing recorded human voices.

In theory the process of voice transformation already discussed can be applied to the output of Festival, but there has been very little work in the area. Festival does have facilities to add sound plug-ins post speech production. The process of voice adaptation is dealt with from the perspective of a TTS system in *Text-to-Speech Voice Adaptation from Sparse Training Data* [Kain et al, 1998]. This paper deals with how it is possible to create a filter to change one voice into another after having as little as 8 sentences worth of source material. It outlines various mathematical formulae used in the process and compares a number of different methods, outlining their complexities, and the quality of the output. The useful thing about this paper is that it was tested and the process developed on the Festival system, which makes it very relevant to my project. The paper goes on to outline the subjective and objective tests

that were carried out to test whether the new voice was in fact any good. The paper states that "Informal perceptual tests reveal that the subjective quality is acceptable, even though the speaker identity of the target has only been partially adapted to." [Kain et al, 1998]. I propose using this paper as a basis for building a tool that allows the easy creation of a filter that will change the basic Festival voice to sound more like the target voice, while only having a limited amount of training material. This is achieved by using an adaptation algorithm, similar to that outlined in the paper, that works on segmental properties such as pitch and spectral characteristics, which relate back to the vocal tract size and shape. It is proposed that there is a regression mapping function based on a Gaussian mixture model to estimate the difference between Festivals voice and the target voice.

Another paper by the same authors *Personalizing a Speech Synthesizer by Voice Adaptation* [Kain et al, 1998] also discusses the process of voice adaptation, but instead of dealing with how the actual process works, this paper looks more at testing the results. It states that a user can personalize a TTS system with a new voice quickly and with ease. "The system transforms source speaker spectra represented by bark-scaled LSFs by means of a probabilistic, locally-linear conversion function based on a GMM. Pitch is adjusted to match target speaker's pitch in average value and variance" [Kain et al, 1998]. The paper concludes that moderate success can be had with a little as one minutes training data, but the quality and individuality of the voice increases as the amount of training data increases. After implementing the previous paper, it will be possible to test these claims and find what the optimum amount of training data is, to get an acceptable output with the minimal data.

6. Final use of the project

The final aim of this project is to find a way of quickly and easily creating voices for Festival, with the bonus outcome of doing so and keeping the database to a relatively small size. The department plans on using this method to create voices for its book reading project. A similar though less ambitious project is discussed in the paper The

Storyteller: Building a Synthetic Character That Tells Stories [Silva et al, 2001]. This paper does go into areas not covered by my project, but the section where the narrator's voice was discussed does give some interesting insight into some of the problems that could be encountered, such as lack of expression to the voice and the need to multiple or at least customisable voices.

7. Conclusion

The process of creating a new voice is long and complex. But by either building a tool set to help automate the process, similar to the InvTool for ModelTalker, or by building a tool that creates filters to adapt Festival's output to the target voice, based on the work done by Kain et al, The process should be made much simpler and easier.

8. References

- Dennis Klatt, *Review of text-to-speech conversion for English*, 1987, J. Acous. Soc. Amer. 82, 737-793
- Microsoft Corporation, *Whistler (Windows Highly Intelligent STochastic taLkER)* <http://research.microsoft.com/srg/ssproject.aspx>, 1998.
- SoftVoice, Inc, *SoftVoice* < http://www.text2speech.com/>
- Bell Laboratories, *Lucent TTS engine* <http://www.lucent.com/press/0199/990128.cob.html>
- AT&T Labs, *Natural Voices* < http://www.research.att.com/projects/tts/>
- Discover Magazine, *The Mathematics of . . . Artificial Speech,* <http://www.discover.com/issues/jan-03/departments/featmath/>, 2003
- Edinburgh Univeristy, *Center for Speech Technology Research*, <http://www.cstr.ed.ac.uk/projects/festival/ >
- OGI School of Science and Engineering, *Center for Spoken Language Understanding* < http://cslu.cse.ogi.edu/>
- University of Delaware, *Speech Research Lab* <http://www.asel.udel.edu/speech/

- Alan W Black, Kevin A. Lenzo, *Building Synthetic Voices* – For FestVox 2.0 Edition, 1999-2003
- Theirry Dutiot, *Short Introduction to Text-to-Speech Synthesis*, Faculté Polytechnique de MonsA, 1996.
- Min Tang, *Voice Transformation*, MIT, 2002.
- Oytun Turk, *New methods for voice conversion*, Bogazici University, 2000
- Patrick Bastien, *Pitch shifting and voice transformation techniques*. TC-Helicon
- Ki Seung Lee, Dae Hee Youn, Il Whan Cha, *A new voice transformation method based on both linear and nonlinear prediction analysis*, Yonsei University, 1995.
- Ben Gillet, *Transforming Voice Quality and Intonation*,University of Endinburgh 2003
- Alexander Kain, Mike Macon, *Text-to-Speech Voice Adaptation from Sparse Training Data*, CSLU, 1998.
- Alexander Kain, Mike Macon, *Personalizing a Speech Synthesizer by Voice Adaptation*, CSLU 1998
- Andre Silva, Marco Vala, Ana Paiva, *The Storyteller: Building a Synthetic Character That Tells Stories*, IST/INESC-ID P-1000-026, 2001.