Creating voices for Festival text to speech system.

M. Hood - g01h0708 - Supervised by: A. Lobb & S. Bangay

4th October 2004

Abstract

Text to speech systems are used extensively in a number of applications, however their usefulness is limited by the number of voices in which they can talk.

Creating a new voice is a complex, time consuming process. This paper looks at the process involved in creating a voice for Festival, an open source concatenative, diphone based text to speech system. It attempts to provide some pointers to help the user through the process.

1 Introduction

The processes involved in the making of a voice for a text to speech system varies greatly depending on what you want to do with the voice, what system you are using and what facilities you have available. This paper gives an outline of the process for creating a voice for Festival, an open source concatenative text to speech system created at the Center for Speech Technology Research at Edinburgh University.

The paper explains the idea of a voice, the diphone lists used to create a voice, recording these diphones, selecting a speaker and recording conditions and finally labelling the recorded files. This paper proposes to help build on these steps and, hopefully, help explain why each step is necessary and give hints to help ease the user through what can be a fairly confusing process when first attempted.

2 Related Work

FestVox is a project undertaken at Edinburgh University which focuses on creating voices for Festival. Its documentation [Black et al, 2003] is still a work in progress but is the best resource for information on voice building for Festival. The basic steps involved in creating a voice for an existing language are outlined in chapter 19 and hence repeating them here will be of little value. Another good

resource worth considering when creating voices is Oregon Universitie's "Centre for Spoken Language Understanding" [http://www.cslu.ogi.edu/tts/]. They have built extra tools, languages and voices for festival, which are all availble from their website.

3 Explanation of a Voice.

A text to speech system takes normal text input and aims to produce spoken words. Over the years there have been a number of approaches taken, each with its own advantages and disadvantages. This paper will focus primarily on concatenative diphone based systems, such as Festival. Concatenative systems work by combining pre-recorded sounds to make up the spoken words. The original examples of these were the infamous talking clocks. In the clocks case, a speaker would record each number and words such as "past", "minutes", "hours" and so on. These words were then combined to tell the time. Such an approach is not feasible for more general text to speech systems, as it is impossible to record every word the input may contain.

The approach taken by Festival is to concatenate diphones, which are phonetic pairs. Any language consists of a number of phones or phonetic sounds. When we talk we merge these phonetic sounds together in distinct ways. It is not enough to just record every phone as the way one of the phones reacts with another varies on a case by case basis. It is not the case that co-articulatory effects depend only on a phones neighbour, but this is a simplifying assumption that means we need only record phone pairs or diphones. A language is essentially a list of all phonetic sounds used to talk the words in that language. Every language has its own list though there are obviously some that will be common to most. When creating a voice it is necessary to record all the diphones for a given language so that any word in that language can be pronounced by the system.

4 Creating the voice

4.1 The Diphone List

Any concatenative text to speech system must have every sound it uses pre-recorded and labelled. It has a fixed choice of sounds to choose from and does not ever mix its own. So when making a voice for a diphone based system one needs to record every diphone pair spoken in the language the voice is intended for. In general this is a square of every phonetic sound used to speak the language.

In reality there is often no need for some of the matches as they are never spoken, and there may, in fact, be extra phonetic sounds to consider. These extra sounds will normally result from strong accents, for example accents such as Scottish or Afrikaans make what are normally soft sounds quite hard e.g. "ch" is given a more "x" sound. It is worth considering adding these extra diphones

to the list if you are making a voice of someone with a distinct accent that you do not want lost.

There is also the problem that if words foreign to the language, like names or borrowed words, need to be spoken by the voice the diphones will be missing and hence the speech will not be able to pronounce them. It is then also important to consider what the voice is likely to be saying and if needed add foreign phonetic sounds to the list, the speaker being recorded may have trouble pronouncing them if they are from another language but at least the result will be better than having none at all.

4.2 Recording the Diphones

There are two options for how you can record the diphones. The first is to speak real words that contain the diphone required. This entails first finding all these real words, but with the help of a linguist this is possible, and then extracting the diphones from the words. Because every word is different, and as a whole word is of no use to the TTS system, it means the labelling will have to be done by hand, and even then the quality is not likely to be that great. This is because it is almost impossible, when speaking real words, to pronounce each one uniformly as the surrounding phones will influence the desired ones.

The other option is to speak nonsense words. It is not enough just to say the diphone, but if it can be encased in other, known, phonetic sounds then it can be pronounce in the context of a word, even if it is not a real word, look at the example at the end of the section. This has the added advantage of making every diphone occur in the same place in the word, which makes labelling much easier later on.

Festival includes basic diphone lists for some languages. In this paper we use the standard US English diphone list produced by Festival, which contains 1396 diphones. This list should suit most people's needs, but can be extended. The format of a diphone list must match the example below or the tools and scripts available to Festival will not be able to use the list. Each line of the file should contain a file id, a prompt, and a diphone name (or list of names if more than one diphone is being extracted from that file).

(us_0001 "pau t aa b aa b aa pau" ("b-aa" "aa-b"))
(us_0002 "pau t aa p aa p aa pau" ("p-aa" "aa-p"))

4.3 The Speaker

It is preferable when making a voice, if possible, to find a trained speaker, or at least someone with experience at talking audibly for long periods. Trained speakers are able to talk in a constant even monotone fashion, this keeps the volume and stress levels the same, which is preferable as the final voice will be of better consistancy. And also due to the number of diphones that need recording (anywhere from 1000 - 2500) it is taxing on the speaker.

It may be worth considering doing the recording over a couple of sessions, but then much care must be taken to ensure the recordings are consistent. This

means setting up the recording conditions the same, ensuring machine settings are constant and that the speaker is in a similar condition i.e. not tired, sick, hung over etc. Some even go as far as recommending that the recordings are done at the same time of day.

4.4 Recording Conditions

The environment in which the recording is done is very important, ideally it should be done in an isolation booth or recording studio. However with modern computer technology it is feasible to record straight onto the computer. All recordings I have made have been done this way in a normal room and the sound quality is acceptable for a basic voice. If the work was recorded in a studio, probably on to DAT, then there is the problem of it being one long recording and not separate files. Oregon Graduate Institute [http://www.cslu.ogi.edu/tts/, 2004] use a system of introducing an identifiable noise between recordings so that they can be easily split once loaded onto the computer. One of the most important features is that the environment is quiet, there should be little or no background noise, and for this reason working at night is preferable.

The choice of microphone is crucial and should be of the highest possible quality. It is widely acknowledged that head mounted microphones are preferable when recording speech. They limit background noise and have the distinct advantage of always being the same distance from the speakers mouth. This makes the sound levels easy to control across sessions and allows the talker to move around, sit etc which will all help as the recording takes a long time. The other option mentioned in "Building Synthetic Voices" [Black et al, 2003] is the use of an electroglottograph, which instead of recording the sound coming out of the speakers mouth, uses electrodes on the speakers throat to pick up vibrations. I have never used one, but they apparently give excellent results [Black et al, 2003].

It is likely, even with trained speekers, that some of the recordings will be unsuitable. It may be possible that you do not need to re-record the sounds, especially if there are only a few mistakes it is tedious to setup the environment and get the speaker back. In these cases one can either grab the sounds from another wave file in the same voice, e.g. the "b" and "p" sounds are very similar, or use a file from another similar voice. Obviously doing either of these will lead to a drop in voice quality, but if there are only a few problem cases and if they are not very common sounds no one will notice.

4.5 Prompting

The diphone list can be used to prompt the speaker before they record the nonsense words. The computer speaks the word to the speaker who then repeats it back for recording. This has the advantage of making the speaker repeat the words in a consistant way and also helps with pronouncing, what can often be, fairly ambiguous sounds.

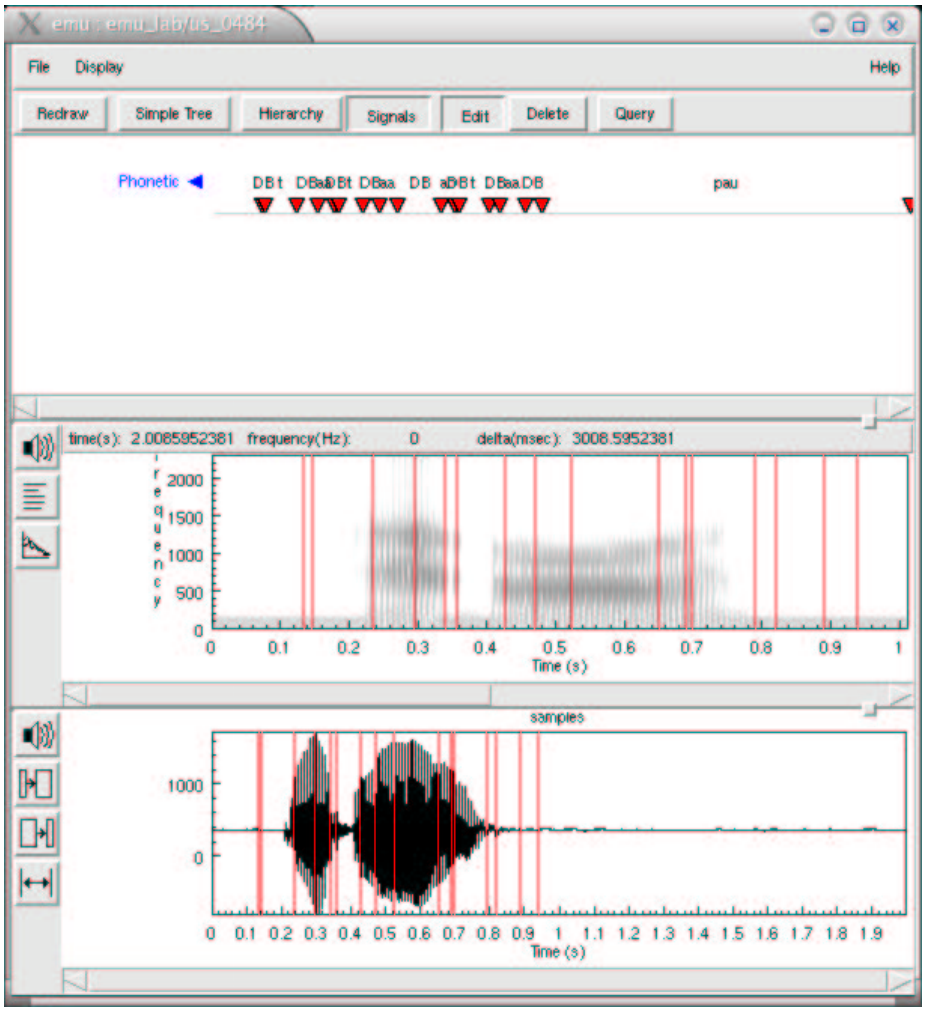
```
1000 ( us_1000 "pau t aa s - ch aa t aa pau" ("s-ch") )
start recording for 2 seconds ...
... end recording
1001 ( us_1001 "pau t aa s - jh aa t aa pau" ("s-jh") )
start recording for 2 seconds ...
... end recording
1002 ( us_1002 "pau t aa s - hh aa t aa pau" ("s-hh") )
start recording for 2 seconds ...
... end recording
1003 ( us_1003 "pau t aa s - th aa t aa pau" ("s-th") )
start recording for 2 seconds ...
... end recording
```

Care must be taken to ensure that the speaker does not start imitating the voice being spoken and loose their own accent. It is of course not possible to do prompts for a new language which does not have a voice to speak in the first place. While one can boot strap diphones from other languages to use in such a situation, I cannot comment on its effectiveness.

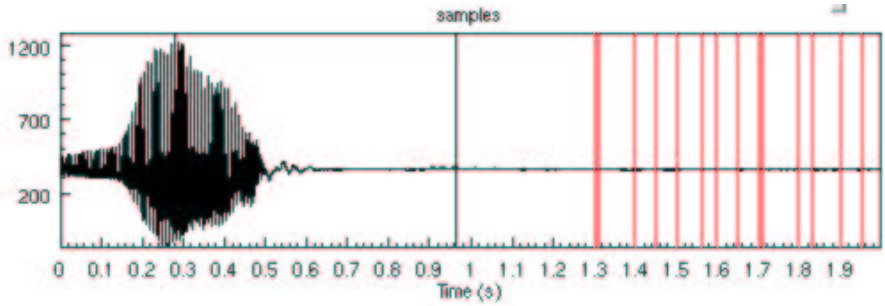
4.6 Labelling

The next step is the vital step of labelling the files. Every *wave* file needs a corresponding label file. The label specifies where the various phonetic sounds are located in the *wave* file, so these can be extracted and played when the TTS needs a specific diphone. The figure over the page shows EmuLabel, a program for working with these labels.

Due to the structure of the nonsense words it makes labelling them much easier than actual words. This also allows for the creation of tools to automate the process. It is possible to hand label all the files, but this is not very practical unless the person doing it is experienced and the final voice quality has to be of a very high standard. The process of labelling is not a quick one, it is repetitive and laborious. The process used in the FestVox scripts is one of marking the boundaries of the phonetic sounds, the actual sound clips are taken from the middle of these boundaries. This gives the labeller a little more leeway when marking the labels as they do not need to be exactly right. It is worth noting that when one is labelling the silence - phone pairs then labelling further into the sound may result in better output, this is due to the fact that the sound level is already dropping into silence if the mid point was taken.



The automatic labeller available with FestVox initially gave impressive results. 22 of the 1396 diphones recorded gave errors when they were used straight after labelling. Theses errors were due to two or more labels having the same values. However on closer manual analysis of the labels it showed that many were labelled completely wrong, some even labelling areas of complete silence.



Generally the errors are clumped together, which implies that certain sounds are difficult to label. This is most likely down to poor recording and in experience of the speaker. The quality of the automatic labelling greatly improves as the speaker becomes more confident with the recording process and pronouncing the nonsense words correctly. Also the better the recording the easier it is for the auto labeller to mark the edges of the phones as they are more defined. Where there were errors in labelling that were not obviously wrong they were, in general, shifted across and not random. This makes correcting them much easier. Black *et al* did an experiment where they used the same voice files and hand labelled one set and auto labelled another. They achieved a reported mean error of 14.77ms and a standard deviation of 17.08ms. My personal experience is that the auto labelling of sound files of moderate to poor quality is much less accurate.

5 Conclusion

When creating a voice there is no alternative to experience. The process is finicky and while I hope this paper has helped to smooth the process, one if likely to find many other pit-falls when making their first voice. It is however possible with perseverance to create a voice whose quality is more than acceptable and very useable even with the most basic of equipment. An example of this is ru_us_matt_diphone, a voice recorded in a normal room, directly onto the pc, using a standard pc micorphone. The diphone list was the normal US English list generated, prompted and labelled by the FestVox tools. EmuLabel was then used to correct mistakes before the final voice was compiled.

References

- Alan W Black, Kevin A. Lenzo, Building Synthetic Voices For FestVox 2.0 Edition, 1999-2003
- Centre for Spoken Language Understanding, http://www.cslu.ogi.edu/tts/