Corpus Collection Protocol: Speech, Song and the Fuzzy Intermediate Domain.

David Gerhard

February 5, 2001

1 Introduction

In this document I will present an outline, specifications and discussion on a proposed protocol for collecting a corpus of sound files containing human utterances. The corpus collection is primarily for my thesis work on fuzzy classification of human utterances on a speech/song axis, but I am planning to build the corpus in such a way that it can be published and used for other research. For that reason I will be making the corpus domain more general and the annotations more informative than perhaps is necessary for my thesis alone.

The building of this corpus will proceed in two stages. Stage 1 is the collection of appropriate sound files from various pre-recorded sources including internet, radio, published sources such as music and spoken word CDs and movie soundtracks, as well as collection from live sources, in the form of solicited utterance samples from human subjects.

Stage 2 of the corpus building procedure is to annotate the corpus. This stage will consist of going through the corpus and transcribing the words, as well as soliciting human subject opinions of the sounds in the corpus. The human opinions will give the corpus validity in the speech/song classification, especially in the fuzzy intermediate domain which will contain utterances such as poetry and chant.

2 Corpus Design

This section presents a discussion on the proposed structure of the corpus as well as the limitations and restrictions that will be applied to the corpus design. A summary of this discussion will be presented at the end of the section.

2.1 Corpus Domain

The FSS (fuzzy speech song) corpus is intended for a specific research domain: human utterance classification in the domain of speech and music. The primary limitations on the corpus are that it will contain only monophonic (with no background or noise) human utterances, containing speech, song, or some intermediate vocalization; and that all samples will be recorded digitally if possible, and stored digitally.

Some secondary restrictions are that the FSS corpus will contain primarily English when a language is used, although the corpus is not restricted to English; samples that contain song will be primarily in the 12-tone equal tempered music system (commonly referred to as the "western" music system) but again the corpus is not formally restricted to the western music system. The corpus will contain a few samples of other languages and other music systems in the corpus for comparison, especially tonal languages and aboriginal music systems.

The corpus will include samples that reflect different characteristics of human speech, as well as different intentions for the corpus itself. The corpus will be able to be segmented along three axes:

- Constrained utterances / Free utterances
- Spoken utterances / Sung utterances
- Speaker Characteristics

2.2 Constrainedness

In order to make the corpus useful for the specific context of fuzzy speech/song classification, but at the same time still be valid for real-world samples, the corpus will contain solicited human utterances of two types. Constrained human utterances will have one or more restrictions placed on the utterance during recording. The proposed constraints fall into four categories:

- Constraints on content of utterance
- Constraints on style of utterance
- Constraints on both content and style
- No Constraints

Content constraints. The constraints on the utterance content consist of requiring the speaker to utter a specific phrase. The phrases to be uttered are chosen to reflect certain expected features of the speech/song classification. Two features that will be investigated in this manner are voiced/unvoiced distribution and formant constancy

It is expected that song will show a higher percentage of voiced segments of speech (vowels, etc.) and a lower percentage of unvoiced phonemes (fricatives, plosives etc.). Indeed, preliminary experiments have shown this to be true. All English lyrical (spoken or sung) utterances contain voiced phonemes, but not all contain fricatives. The voiced/unvoiced distribution feature extractor would believe that all utterances with no unvoiced phonemes are song. For this reason, the corpus should contain a spoken utterance with only voiced phonemes to make sure the full system can handle such an utterance. It is proposed that one of the spoken utterances solicited from subjects be:

"When you're worried, will you run away?"

Another feature expected in song is that the glide of diphthongs will be suppressed till the beginning or ending of the phoneme. To test this, it is desired to have utterances with many diphthongs. For this reason, it is proposed that one of the utterances solicited from subjects be:

"Row, row, row your boat, gently down the stream."

The diphthongs in this utterance are expected to be short and rhythmic. As a contrast, the following utterance will also be solicited:

"O Canada, our home and native land."

Both of the above utterances will be solicited spoken as well as sung.

Style constraints. Because the work is in essence an attempt to distinguish between speech and song, with investigations also directed toward intermediate vocalizations, Part of the corpus will contain utterances where the subject is prompted to sing or is prompted to speak. As indicated above, some samples will be requested in both spoken and sung styles, so the differences between speaking and singing in these samples would not be obscured by differences in content or in subject characteristics.

There would be samples taken of unconstrained content with constrained style as well. The purpose of these samples would be to expand the corpus beyond constrained utterances, which test particular characteristics and features, but are not appropriate for design of a system to operate on "real-world" data.

The style-constrained samples would allow the speaker to choose the content (lyrics) of the utterance, but would insist on a particular style of utterance. Example prompts are:

"Sing the first line of your favorite song."

"What did you have for lunch yesterday?"

A further style constraint which will attempt to illicit samples in the middle ground between speaking and singing would allow the speaker to utter any lyric in any style so long as it is *not* speaking or singing. A prompt for this style constraint would be:

> "Tell me what you did last weekend, using words, but without speaking or singing."

A similar prompt using constrained content would be:

> "Utter the phrase 'Why is the sky blue?' without speaking or singing."

The phrase "Why is the sky blue?" has many characteristics that are desirable for this corpus. It contains a good distribution of fricatives, both voiced and unvoiced, and two diphthongs which rhyme.

No constraints. This section of the corpus will consist of samples that are unconstrained in any way. These samples include all "found" samples (samples not directly solicited from human subjects), for example samples taken from radio or from movie soundtracks. The corpus will also include some unconstrained samples solicited from subjects. The majority of the corpus that I currently have falls into this category. The richness and variability of completely free samples fills out the structured nature of the rest of the corpus.

2.3 Utterance class

The second way of dividing the corpus is in the perceived style of the utterance itself. Since the majority of the research is concentrating on speech and song, many of the samples will fall clearly into one of these two categories, with the remainder falling into the category of "Fuzzy speech/song", indicating that the sample has characteristics of both speech and song, but is not clearly one or the other. Some samples will be specifically designed to fall within this category, such as the manipulated sample corpus described in Section 3.4, as well as some of the style-constrained utterances. Some found utterances will end up in this category as well. The possible utterance classes are:

- Purely speech utterances
- Purely song utterance
- Fuzzy speech/song

It is important to note that this classification will rely on human opinion testing of the corpus, and not from any characteristics of the corpus files. The entire corpus, once collected, will be labeled on a fuzzy scale between speech and song, using the results of the human opinion testing described in Section 4. It is expected that there will be many samples characterized as pure speech or pure song, which is why the corpus collection protocol is biased toward samples which are expected to fall into the fuzzy category between speech and song.

2.4 Speaker characteristics

Human speech is varied because human speakers are varied. Since the purpose of the proposed corpus is to aid in the design and testing of a system that will operate on human speech, it is important that the corpus contain a balance of human speaker characteristics. For this reason, it will be important to make sure that the subject base contains a good balance of individuals on the basis of the following characteristics:

- Age
- Gender
- Musical/Speech training

Young people, especially children, speak with higher pitch than do adults so the pitch range feature extractor proposed in the classification system should be able to handle speech from children. Older people have different voice characteristics, as do children at the verge of puberty. The proposed corpus would do well to have samples from representatives of each of these age groups.

Men and women have different pitch aspects of speech. The proposed corpus will have a balance of male and female subjects.

A characteristic of speech that is especially relevant for this corpus is musical training. Song is a faculty that all humans possess, but those that are trained in singing have the ability to make their voice do exactly what they want. These speakers will be able to give samples of very high quality song, and might be more able to give samples in the middle-ground between speech and song, or samples that are neither speech nor song. Individuals who are trained in speech, such as actors or radio personalities, also have the ability to manipulate their voices as desired. The corpus should have a portion of samples solicited from trained users of speech and song.

3 Corpus Collection

This section describes the protocol for collecting the samples which will populate the FSS corpus as described above. There will be four categories of collection:

- Free samples
- Constrained Samples
- Found Samples
- Manipulated Samples

Each category is described here, including proposed collection protocol. The solicited samples will be collected from human subjects using a protocol approved by Simon Fraser University according to the university research ethics guidelines.

The subjects will be selected randomly, with intention to fill out the categories described in Section 2.4.

3.1 Free Sample Subcorpus

An important sub-corpus is the corpus of solicited samples with no constraints. As discussed above, these samples are necessary to fill out the otherwise structured nature of the corpus, and also provides some "real-world" samples for a system designed on this corpus to deal with.

The proposed collection protocol for unconstrained samples is this: Two unconstrained sample phrases will be collected from each subject, using the following prompt for both samples:

> "Please speak or sing anything you like for about 5 seconds."

3.2 Constrained Sample Subcorpus

This subcorpus will be gathered from human subjects, in the same way that the free sample subcorpus will be collected, using various constraints as described in Section 2.2. The samples will be constrained in style, in content or in both style and content. The proposed prompts, as stated above, are:

> "Sing the first line of your favorite song."
>
> "What did you have for lunch yesterday?"
>
> "Please speak the phrase 'When you're worried, will you run away?' "
>
> "Please sing the phrase 'Row, row, row your boat, gently down the stream.' "
>
> "Please speak the phrase 'Row, row, row your boat, gently down the stream.' "
>
> "Please sing the phrase 'O Canada, our home and native land.' "
>
> "Please speak the phrase 'O Canada, our home and native land.' "
>
> "Utter the phrase 'Why is the sky blue?' without speaking or singing."
>
> "Tell me what you did last weekend, using words, but without speaking or singing."

As with the unconstrained samples, the subjects will be encouraged to limit their utterances to about 5 seconds. For prompts that require a specific phrase, the user will be encouraged to read, remember, then speak the phrase as if they were talking or singing to another human. For prompts requesting an utterance that is neither speech nor song, the subject will be encouraged to practice a couple times before recording, to get a feel for what a non-speech, non-song sound might be like.

3.3 Found Sample Subcorpus

This subcorpus will be populated by extracting short segments of sound from publicly available audio, such as radio, published music, movie soundtracks, and .wav and .mp3 files available on the internet. Copyright laws allow reproduction of copyright material for research purposes.

I will be scouring the net, radio and movies for sounds that would be appropriate for this corpus. Examples of sounds that I am expecting to acquire:

- "Daisy, daisy, give me your answer, do" (HAL 9000, "2001")
- "Good morning vietnam!" (Robin Williams, "Good Morning Vietnam")

Various vocalists have been suggested to me as well including Mark Knopfler, Yoko Ono and Bob Dylan. The challenge with collecting found samples will be to find samples of people singing and speaking without any background noise or music. Stationery background noise is acceptable, because the system will be able to filter it out as long as there is a couple seconds of silence (with the background noise) before the human utterance begins.

3.4 Manipulated Sample Subcorpus

This subcorpus will consist of samples that have been deliberately manipulated to fool a specific feature extractor. Three of the feature extractors that are expected to be successful in detecting the presence of song are:

- Pitch outside of normal pitch range.
- Presence of vibrato in pitch track.
- Larger proportion of voiced segments.

To design sounds that would fool each individual feature detector, I would begin with a sound that would clearly be classified as speech, and then manipulate characteristics of the sound using granular synthesis. The goal of this manipulation would be to create a sound that a human would consider to be speech, but which has one of the features of song as expected from the feature extractor being designed.

As an example, I would take a sound sample of someone speaking, with pitch inside the normal pitch range for speaking, and granularly increase the pitch so that the pitch range feature extractor would classify it clearly as song. Opinions of this file would be solicited in the usual manner (see Section 4) to determine what effect the pitch range has on the perceived class of the sound.

This procedure would be repeated with the other features: adding a harmonic ripple to voiced segments of a speech sound; extending the voiced segments and compressing the unvoiced segments; and performing similar manipulations with other features.

The same procedure would be performed in the other direction - making song samples sound like speech for a particular feature extractor. For example,

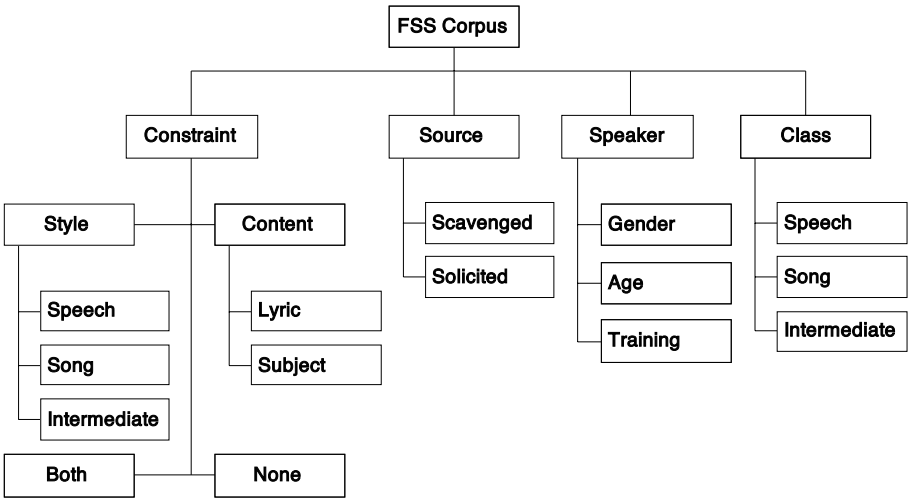

Figure 1: The FSS Corpus subdivisions and categories.

Table 1: FSS Collected Subcorpora characteristics

Subcorpus	Collection Method	Size	Purpose
Free	Solicitation	100	"real-world" samples
Constrained	Solicitation	600	boundary conditions
Found	Extraction	100	existing samples
Manipulated	Design	50	individual feature testing

removing spectral ripple from a song sample; bringing the pitch into normal speaking range; compressing the voiced segments and extending the unvoiced segments; and performing similar manipulations with other features.

The goal of this sub-corpus would be to verify the individual success or failure of each feature for a speech/song classification, as well as testing the robustness of the overall system in the presence of one divergent feature result.

3.5 Corpus Summary

Figure 1 shows the proposed corpus by subdivision criteria. Table 1 summarizes the proposed subcorpora by collection procedure, along with the related collection methods, expected sizes, and purposes.

Each sample in the corpus can be classified on each axis. if a sample were solicited under the constraint of style = song, for example, the source would be "solicited" with the gender, age and training characteristics corresponding to the subject, and the class would be indicated by the opinion gathering after the corpus is fully collected.

4 Opinion Solicitation

The second stage of building the corpus is to annotate the corpus. This consists of transcribing all lyrics used in the speech and song samples, as well as soliciting human opinion scores for all samples in order to label the corpus on the "speech/song" axis. The opinions will be solicited in a manner similar to the solicitation of the samples, and opinions will be solicited from the subjects who provided the samples, as well as other subjects who did not.

4.1 Opinions on the Full FSS Corpus

Depending on the size of the corpus, subjects will be asked to provide an opinion for some or for all of the corpus. 850 samples of 5 seconds each would take 1 hour, 11 minutes to listen to, without pauses between samples. If we predict 15 seconds for each sample, listening and classifying, the time to complete 850 samples would be 3 hours, 33 minutes.

The opinions will be recorded with the age, gender and musical or speech training level of the subject, along with whether or not the subject provided samples for the corpus in the corpus collection stage.

The subjects will be asked the following questions about each sample:

"Please rate this sample on a scale between speaking and singing."

"Please rate the quality of the speech or song, from 0 to 10."

"Please write one or two words to describe this sample."

4.2 Opinions on a Selected Sub-set of the FSS Corpus

Along with these general opinions, a small sub-set of the corpus will be selected for further opinion gathering. This subset will consist of solicited, found and designed samples which fall into the following categories:

- Clearly speech
- Clearly song
- "Rap" style utterance
- Poetry
- Chant
- Monotonous speech (as in a university lecture)

Also in this sub-set will be samples from the corpus that are difficult to categorize, or perhaps samples that fall into the fuzzy middle ground between speech and song.

More detailed opinions on this sub-corpus will be requested. The subjects will be asked the following questions about each sample:

"Please rate this sample on a scale between speaking and singing."

"Please rate the quality of speech or song, from 0 to 10."

"Please indicate what the speaker might have done to make this utterance more speech-like"

"Please indicate what the speaker might have done to make this utterance more song-like"

The first two questions are identical to the questions for the full corpus, and are included in the sub-corpus opinion testing to test for opinion consistency. The second two questions are free-response, and are included to extract a general intuition about speech, song, and the middle-ground between them.

5 Summary

This document describes the proposed protocol for collecting and annotating the FSS (Fuzzy Speech Song) corpus, intended for research on human utterance classification, specifically speech, song and the fuzzy intermediate domain between speech and song.

Stage 1 of the corpus collection protocol consists of acquiring utterance samples from human subjects and from available media, in four categories: Constrained utterances, Unconstrained utterances, Found utterances, and Designed utterances.

Stage 2 of the corpus collection protocol is the annotation of the corpus by human subject opinion. Human subjects will be asked to listen to the samples in the corpus and provide opinions based on a series of questions. Subjects will also be asked to provide more specific opinions on a subset of the corpus.