SR4X Verification May 23, 1996

1 What you'll hear and why

The speech you are about to listen to is intended to be used for research in speaker recognition. Speaker recognition enables computers to tell the difference between people's voices. When you speak, the system should be able to distinguish your voice from someone else's voice.

One difficulty with speaker recognition is that our voices sound different depending on the phone we are calling from, the time of day, our mood, etc. That is why we designed the data collection to be done over time (four separate calls) with different phones (four channel types) in different locations (four different calling locations). With this database we have recorded each given voice in a variety of situations.

2 The most important thing

As you listen, pay close attention to who is talking, what they say, and what kind of phone they are using.

3 Diagram of a filename

Can you predict the identity of the caller, what they are to say, and the channel type? Yes, simply by looking at the filename. Here is an example filename with the significant parts labeled:

```
SD-7876-tektronix-t3-20.wav
   ^     ^     ^
   |     |     |
   a     b     c
```

(a) This is the call number. The call number uniquely identifies each caller. Watch the call number closely. Every file with the same call number should have the same speaker. If the voice changes when the call number has stayed the same, mark the call as bad.

(b) Next there is a dash followed by the question type. The question type is what the person is supposed to say. For this file, as you can see, the person was supposed to say "tektronix."

(c) After the question type is the channel (phone) type.

- t1 = office phone (internal line)
- t2 = home (external line)
- t3 = carbon microphone
- t4 = speaker phone

See section 4.3 for more detail.

4 Classification

You will be asked to classify each file into one of four different categories. See Table 4.

- Bad
- Different
- Noisy
- Good

4.1 Bad

A file is bad if the person says the wrong utterance. For example, the speaker says *generations* when asked to say *sangita*. If the word the caller says is just slightly different from the word we asked them to say (generations for generation) the utterance should be coded as different not as bad.

A file is also bad if the wrong speaker is responding. Occasionally two different people will have the same call number, so you must be careful to note if the call number changes at the same time the voice changes, and vice-versa.

Finally, a file should be marked as bad if the noise level is so high, or the accent so bad that either the word spoken or the identity of the caller could not be reliably determined. Think of this as a competition in which the goal

Table 1: Table 1: Summary of classification scheme

BAD	*DIFFERENT*
1. totally wrong word 2. wrong speaker 3. so noisy you can't fairly determine whose speaking or what they are saying	1. "phonetic" changes including: a. addition or deletion of phonemes or a syllable b. addition of a syllable or a word 2. very strong foreign accent 3. extra human, non-speech sounds (clear throats, breaths, etc) 4. background talking
NOISY	*GOOD*
1. loud line noise relative to session and channel type 2. long pause at end or beginning of file 4. keyboard sounds	1. right word 2. right speaker 3. not noisy 4. not different 5. not bad

is for a human to correctly identify what was said and who said it. If the utterance is clear enough that it would be *fair* to ask someone to identify it, then it is not bad. But if it is *unfair* to expect someone to reliably identify the speaker or the utterance, the file should be classified as bad.

4.2 Different

A file should be classified as different if the actual pronunciation (phonetic realization) deviates from the norm in a significant way. Significant deviations include

- a very strong non native accent
- the addition of one or two phonemes, up to the addition of a single syllable word: **uhm generation** or **generations**
- the deletion of a phoneme or at the most, a syllable: **sangit***. If the deletions or insertions are larger than that listed here, the file should be coded as bad.

A file should also be coded as different if there is background talking (speech that is audible, not spoken by the caller).

And lastly, a file is different if there are any non speech events such as clear throat, cough, laugh, breath, tongue click, or any other sound made by the caller that is neither speech nor background noise.

4.3 Noisy

Excessively noisy calls should be coded as noisy. "Excessive" is admittedly subjective, but our goal is to remove as much of the subjectivity as possible.

If the following noises occur, the file is in the *noisy* category:

- clearly audible keyboard, paper shuffling
- loud line noise relative to channel type (A few clicks are okay, and are not to be coded as noisy.)
- loud background noise
- Long pauses at the end of the file

Something else to keep in mind is the four different types of phones from which participants will be calling:

1. **Speaker phone** This tends to sound distant or echoey at times. Files should not be coded as noisy for echo sounds or for a distant quality when the caller is using a speaker phone.
2. **Carbon Microphone** The calls from the carbon microphone call will normally have a *tin can effect* so the tinny sound is not to be classified as noisy.
3. **Internal Line** Internal line calls should be quiet, normally. Mark noisy if it has a lot of line noise or background hum, or other noise.
4. **External Line** Same coding as for an internal line. Often there is more background noise (washing dishes, music) and it should be coded as noisy if you hear these types of noises.

4.4 Good

After defining the previous categories, *good* is easy to define. A good file is one in which the correct person says the correct utterance with little or no line or background noise.

4.5 Examples with coding

- "generations" for "generation" **different**
- "hawaii" for "sangita" **bad**
- noisy call compared to rest by caller with same channel **noisy**
- "generation ohh" **different**
- "generation can you get me a cup of coffee" **bad**