IT608 Course Project:
Profile based speech coding using clustering on elementary sounds

Chetan Vaity (02329901)
Shweta Agrawal (02329013)
Amreek Singh (02329025)

21 April, 2003

1 Objective

This work is based on the intuition that in a speech sample of a particular person, similar *elementary sounds* are repeated. For example, when a person says "cricket" and "club", the initial "kk" sound in both words will have similar characteristics. Significant reduction in storage could result if the actual signal information for both these sounds is not stored. Instead the *elementary sound* is stored just once and wherever this sound appears the same stored sound is played.

The purpose of the current project is to demonstrate the feasibility of this idea.

2 Design

Consider the audio news downloads which appear on news websites. These news items are typically read out by one person (or a small group of persons). The speech *profile* of this person can be created which will contain the collection of *elementary sounds* uttered by him. This profile will be a one-time download for the listeners. The actual news audio samples can be encoded based on the profile. The users will only need to download the encoded data (which will be much smaller than the actual audio data). This can be decoded using the profile stored earlier by the user, and the audio can be regenerated.

The steps followed in the current project are detailed below.

2.1 Profile generation

A recorded lecture was obtained from the DEP (Distance Education Programme, IIT Bombay) archives. All experiments have been conducted using this 2 hour sample (sampling rate: 22050

Hz, single channel).

A 15 minute sample was extracted for profile generation. This file was divided into 45000 files of 20 ms duration each, using the `sox` utility [2]. 10 MFCC features (Mel-frequency cepstral coefficients) [1] were computed for each of these *sound-slices*. MFCC features are perception based features which are widely used in the speech recognition arena. This was done using a software called `cepstral` [4].

It was assumed that 10000 different *elementary sounds* will be enough to characterize the range of sounds produced by a person. This number was arrived at empirically. The 45000 sound samples were then clustered into 10000 clusters based on their MFCC features. A bisection variant of the k-mean algorithm was used for clustering. The `vcluster` tool of the CLUTO [3] clustering suite was used for this purpose. For each of these clusters, a representative sample was chosen randomly. A better approach can be followed in future work.

These 10000 representative sound samples are then assigned unique codes (the cluster numbers have been used as the codes). This collection of representative sounds and their codes is the profile, using which other sound samples can now be encoded.

2.2 Encoding

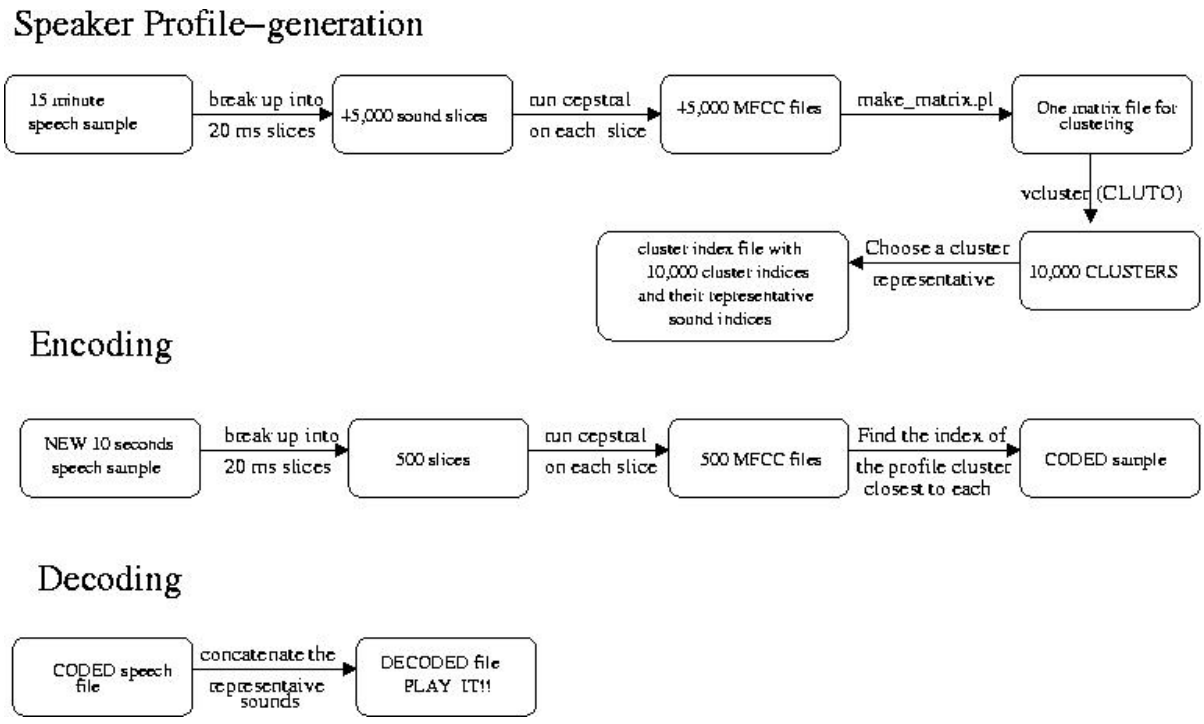
A new 10 second sample was taken from the lecture (disjoint from the earlier sample, of course) and divided into 20 ms slices. MFCC features were extracted from the 500 sound-slices created this way.

Each of these feature vectors was taken and a closest match was found from the 10000 feature vectors of the representative samples of the profile. This was done by determining the minimum euclidean distance in the 10 dimensional feature space.

Thus, for each of the 500 sound-slices, a representative sound from the profile was identified. The encoded file consists of this sequence of codes of the representative sound samples.

2.3 Decoding

The decoding is done using the encoded file and the profile (i.e 10000 representative sound-slices). The resultant audio is created by successively concatenating the representative sound samples indicated in the encoded file. No smoothing was done due to time constraints, and it is believed that if done, smoothing will improve the quality of the resulting decoded sample.



3 Results and Conclusion

The sound of the decoded sample is discontinuous but intelligible. Obtaining recognizable speech with such naive techniques is extremely encouraging.

The ratio of the sizes of the encoded file to the sound sample is around 200. This can be further increased using direct binary data in the encoded file instead of the ASCII text which was used.

4 Issues

- A better MFCC feature extraction program should be used for further work, as the current one (`cepstral`) is not open source and is not robust.
- The method of physically dividing the sound file produces a large number of small files which tax the operating system resources. A better method would be to extract features from particular regions in the sound sample file.
- The empirically arrived at figure of 10000 for the number of representative sounds needs to be further researched.
- Currently 10 MFCC features are being used for sound characterization and clustering. Preliminary experiments had shown better results with just two of the features (7th and 8th MFCC features). It is possible that other features or a subset of these features give better results. This needs to be investigated.

- Techniques exist for smoothing of sounds generated by concatenation. These should be applied to improve the audio quality.
- **Segmentation**: Initially, segmentation of the speech sample was tried using a statistical technique. This technique is based on determining the boundaries for segmentation using discontinuities in the MFCC feature trajectory. [5]. This approach was abandoned because it produced unequal sized sound samples which were difficult to compare and replace.

5 Effort details

- Percentage work done by each member: 33% each
- Total time spent: approximately 80 man(/woman)-hours

References

[1] Suresh Balakrishna, Speech Recognition using Mel Cepstrum features, Mississippi State University, 1998

[2] SoX - Sound Exchange (http://sox.sourceforge.net)

[3] CLUTO, A clustering toolkit (http://www-users.cs.umn.edu/ karypis/cluto/)

[4] cepstral, MFCC feature extraction algorithms (http://tiger.la.asu.edu/software.htm)

[5] George Tzanetakis and Perry Cook, Multifeature audio segmentation for browsing and annotation, *Proceedings 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999*

[6] Regine Andre-Obrecht, A new statistical approach for automatic sound segmentation of continuous speech signals, *IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 36, No. 1, January 1988*

[7] ISIP Automatic Speech Recognition (http://isip.msstate.edu/projects/speech/)

[8] An online Automatic Speech Recognition course (http://isl.ira.uka.de/speechCourse/overview/contents.

[9] John Watkinson, The Art of Digital Audio, 3rd Edition