

# Language Independent Speech Compression

Lakshmi Gopalakrishnan, Virag Shah, M. Habibullah Pagarkar, Nimish Sheth, Rizwana Shaikh

## Abstract

This paper proposes a method to develop a complete speech compression system using Devanagari script. This involves developing a phonetic recognizer which segments the words into a set of phoneme units (diphones and consonant clusters) available in Devnagari Script. This is done using Hidden Markov Models (Baum-Welch and Viterbi methods). These recognized units are mapped onto a labeled phonetic database consisting of 850 phoneme units. The uttered speech can, thereby, be expressed using only the pointers to these entries. This results in speech compression. Weights are then assigned to each phoneme, which specify the pitch, the amount of stress laid while pronouncing it and the speed with which a person utters it. The speech is then reconstructed by applying concatenation synthesis for smoothing at segment boundaries. The use of Devanagari, here, has a distinct advantage over other languages since it is phonetic in nature.

## Introduction

The aim of speech compression is to produce a compact representation of speech sounds such that when reconstructed it is perceived to be close to the original. The two main measures of closeness are intelligibility and naturalness. There have been different approaches towards speech compression where speech is treated as yet another type of signal; redundancies are identified and eliminated, thus effecting compression.

We sample the signal at regular time intervals and quantize the value obtained. Majority of the losses during speech compression occur during the quantization phase. For recorded speech to be understood by humans, an 8 KHz sampling rate or more and at least 8 bit sampling is required. This produces poor quality, but understandable speech. If we group the samples into blocks called frames, we get a better representation for the signal making it possible to analyze the entire frame. Within each frame a vector of features is stored. The final step is vector quantization. Improvements can be achieved by increasing the number of bits in sampling to 12 bits or 16 bits, or by using a non-linear encoding technique such as  $\mu$ -law or A-law.

However  $\mu$ -law coding does not exploit the sample to sample correlations found in speech. ADPCM is the next family of speech coding techniques, and does exploit this redundancy by using a simple linear filter to predict the next sample of speech. The resulting prediction error is typically quantized to 4 bits thus giving a bit rate of 32 Kbps

Our approach is based on the very nature of speech where we are attempting to establish a representation for speech in terms of subunits. This representation encapsulates knowledge on multiple linguistic levels including morphology, syllabification, stress, phonemics and graphemes. This new paradigm can well be used to more efficiently model the words in a language. The information extracted can be utilized in a variety of different speech applications, thus providing enhanced performance.

At least two immediate applications of this representation exist. First, word can be composed from the set of these finite units, much like a function can be composed from a basis set. A speech recognizer could operate with these underlying sub-word units, leading to unlimited vocabulary recognition. Second, this theorized "alphabet" could also be used in letter-to-sound/sound-to-letter generation. Once the correct sequence of these units is found for a word, the phonological information could be directly inferred from these units.

A fixed dictionary of phonemes is maintained at both the sender and the receiver. Once we know the possible phoneme, we characterize it and pick out features that will enable us to properly reconstruct it at the other end. Next a data structure corresponding to every phoneme is constructed along with the characteristics and is then sent to the other end where it is uniquely decoded and we get our phoneme. Using concatenation synthesis, we can now reconstruct the entire sound sequence.

Devnagari is the language of choice over the more wide-spread English for it is an inherently phonetic language and is spoken the way it is written. All human languages use a limited repertoire of about 40 to 50 sounds called phones. Combination of characters in different contexts produces different phones. We believe that any spoken language can be represented as basic phonemes in Devnagari along with a few additional phonemes particular to that particular language. For e.g. Australian English is heavily accented and Russian is very guttural. These introduce new phonemes into consideration. However, we will be first working on the most basic of phonemes and plan to gradually increase the number of phonemes and thus the languages which can be transmitted.

There are many difficulties in this method. Firstly, the existence of homophones or similar sounding words confounds the problem. Another problem is deciding where the phoneme ends. This is known as the problem of segmentation. For example, when listening to a foreign language, it seems that all the words are continuous without a break in between. Actually, words are continuous even though they are written separately and thus we need very good techniques to identify the word units.