Effects of Audio Compression on Speech Recognition

Introduction

Speech recognition has become widely popular and it is very common to be used as an add-on for various applications, often mobile that use wireless communications. Although certain speech recognition tasks can be performed by the device itself, complex cases usually require lots of memory and power so speech recognition servers are used. In this paper we'll focus on the client-server approach which adds some constraints to the amount of data that can be send over the networks.

If the client application can be complex, certain systems rely on the client extracting the necessary acoustic features of speech, thus compressing and sending only the relevant data. In certain cases the client has to be simple and cannot extract such information, so the speech signal has to be transmitted to the server. Because of bandwidth and delay limits, the signal usually has to be compressed which somewhat alters the original signal.

In the next sections we'll look at these two system approaches and analyze what methods exist to prevent compression from reducing the quality of speech recognition.

Speech Signal Compression

First lets analyze the effects that compression can have on speech recognition systems, which can tells us if there is an actual problem when merging compression and recognition.

In [1], the authors used samples of compressed speech signals (using standard methods GSM, MPEG and G7XX) on a French speech recognition system. The results show a relevant degradation of accuracy due to the coded data.

…

Acoustic Features Compression

As mentioned above, compressing the entire audio signal can lead to lower speech recognition accuracy as relevant speech properties are lost on decompression. In some cases it is not possible to adjust the speech recognition system to compensate for this loss, so we need an alternative.

In cases where the client is expected to have a certain degree of complexity, it is possible to envision a speech recognition system where the client collects the speech data, extracts speech relevant features (MFCCs or similar), compresses the features and transmits the encoded data to the server. The server in turn will receive said data, decompress it and use it as input for the speech recognition engine. In this system the compression happens after extraction so, as long as the compression of the features has acceptable noise, the relevant information is kept intact.

Ramaswamy and Gopalakrishnan in [2] describe a compression algorithm specific for encoding acoustic features.

Conclusion

…

References

[1]. L. Besacier, C. Bergamini, D. Vaufreydaz, E. Castelli. THE EFFECT OF SPEECH AND AUDIO COMPRESSION ON SPEECH RECOGNITION PERFORMANCE. IEEE Multimedia Signal Processing Workshop, Oct 2001, Cannes, France. pp. 301-306, 2001.

[2]. G.N. Ramaswamy, P.S. Gopalakrishnan. COMPRESSION OF ACOUSTIC FEATURES FOR SPEECH RECOGNITION IN NETWORK ENVIRONMENTS. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1998.