

Phonytalk™

Background

Considerable effort has been devoted to designing test signals that have speech-like properties to better evaluate the performance of network equipment. Although single frequency tones are effective for simple, linear analogue network elements, they are inappropriate for testing the complex devices that have become widespread in the last twenty years. The different approaches to the test signal problem have resulted mainly in a number of signal processes being bolted together, for example, spectral shaping, noise, temporal envelope structure, variable power spectral density, probabilistic amplitude distribution etc. These rather complicated signals that were produced still failed to stimulate non-linear devices, such as low bit rate codecs, in a speech-like way.

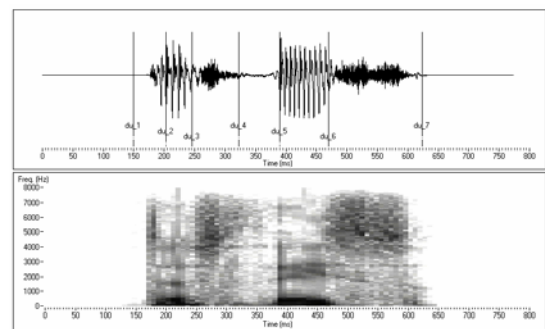
Markov Spherically Invariant Random Processes (M-SIRP) signals were perhaps the most sophisticated test signals devised to date. However they are inadequate due to use of short uniform segments which do not reproduce all required speech properties and introduce non-speech properties at segment boundaries. The best signal to test the effectiveness of a speech transmission path must be speech - a Phonytalk(TM) test signal contains examples of many different properties of speech delivered in convenient package.

The Derivation of the Artificial Speech Test Signal

Speech consists of a certain number of sounds that are capable of being phonemically grouped. Each phoneme is subject to a wide speaker dependent variation and will be affected by its association with other phonemes and the context in which it is produced. The best way to test the effectiveness of a speech transmission path is with real speech. A large number of speakers and speech material is required for the test to be sufficiently rigorous in order to represent all naturally occurring speech. Many sounds will be repeated which represents redundancy in a test stimulus for objective measurements.

One aspect of speech transmission that is affected by the presence of non-linear processes is speech level. Speech level errors lead to quality degradation. Low bit rate codecs can produce erroneous results with the same sound at different speech levels. The use of tones and complex test signals will not necessarily produce a result representative of in-service operation. The best test signal is real speech. Unfortunately the very large number of different sounds in natural language renders the test process unwieldy. A representative subset of sounds would make the process more efficient.

A large corpus of conversational speech material was phonemically transcribed and groups of phonemes categorised according to acoustic characteristics. The relative frequency of occurrence, transitional probabilities, etc were extracted from the corpus. The acoustic groupings might provide for types of vowel sounds subdivided into front, middle, back, round, short, long and so on. The speech sounds can be formed into linguistically legal sequences that include the transitional probability results. These sequences, which are statistically representative of thousands of other related sounds, last for about 25 seconds. This phonemic string can be applied to a speech synthesiser to produce a sequence of sounds, which can be described as a linguistically motivated test signal. The Laureate Speech Synthesiser utilises a very large number of very short speech sounds in a database. Diphone and triphone concatenation is used to provide the closest possible match between the symbol stream and the collection of speech sounds used by the synthesiser to produce the desired sounds. The linguistically motivated test signal, produced by Laureate, has been transcribed by phoneticians to confirm that it matches the desired symbol stream. The resulting signal is thus representative of the full range of sounds and transitions that occur in natural conversational speech, and yet is of a practical duration for testing. British and American English speech samples are immediately available. Other language versions will follow.



Label	Freq. Occurrence	Previous Trans.	Following Trans.
du_1	0.0450	0.266	0.078
du_2	0.1009	0.035	0.037
du_3	0.0506	0.073	0.001
du_4	0.0289	0.001	0.052
du_5	0.0387	0.038	0.010
du_6	0.0506	0.007	0.361