Introduction

- The VODER "Voice Operation DEmonstratoR" of Homer Dudley, demonstrated at Bell Laboratory exhibit at the 1939 New York World's Fair, was controlled using a keyboard and foot pedals.
- We can say that these peripherals enabled to control parameters of the a vocoder behind the VODER. And operator of the VODER was a "model" that generated the control sequence.
- In the case of the VODER the "model" to synthesize the speech parameters was a human. Current vocoders incorporate the modelling of the parameters. To distinguish them from historical vocoders, we are going to call them hereinafter *synthesis vocoders*.

Analysis - MGC features

Mel-generalised cepstral (MGC) features $c_{\alpha,\gamma}(m)$ are typically used in speech vocoding.

$$H(z) = s_{\gamma}^{-1}\left(\sum_{m=0}^{M} c_{\alpha,\gamma}(m) z^{-m}\right) = \begin{cases} \left(1 + \gamma \sum_{m=1}^{M} c_{\alpha,\gamma}(m) \tilde{z}^{-m}\right)^{1/\gamma}, & -1 \leq \gamma < 0 \\ \exp \sum_{m=1}^{M} c_{\alpha,\gamma}(m) \tilde{z}^{-m}, & \gamma = 0 \end{cases} \tag{1}$$

where M is an analysis order.

Relation of α and γ

- The variable $\tilde{z}^{-1}$ can be expressed as the first order all-pass function

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{2}$$

  where α is a warping factor.
- For 16kHz, $\alpha = 0.42$ gives good approximation to the mel scale. The parameter γ control the representation accuracy of poles and zeros.
- As the value of γ approaches zero, the accuracy for spectral zeros increases at the expense of formant accuracy.

Relation of MGC to other analysis methods.

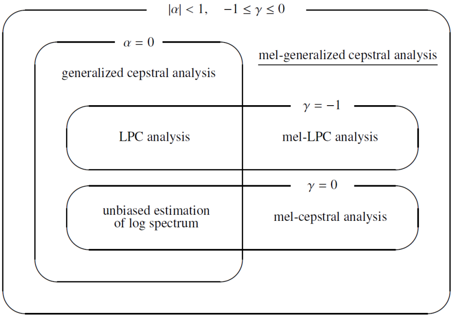

Figure: Relation of MGC to other analysis methods.

For more details and explanation, please see Phil's root cepstrum notes.

Speech parameter generation

- As already mentioned, vocoders enable to model their parameters. The models are typically Hidden Markov Models (HMMs).
- Then, an additional algorithm need to be used, to calculate the speech parameters (static cepstra) from continuous mixture HMMs with dynamic features.
- An iterative MLPG algorithm does it. We will not explain it as it is a staff for sequential speech processing systems.

Re-synthesis

- In mel-generalised speech ceptrum $H(z)$ is modelled by as set cepstrum coefficients $c_{\alpha,\gamma}(m)$.
- For re-synthesis, the parameter γ is fixed to be $-1/2$. This value balances good representation of both spectral poles and zeros.
- Then, the synthesis filter is realised as a rational transfer function

$$H(z) = \frac{1}{\{B(\tilde{z})\}^2} \tag{3}$$

where

$$B(\tilde{z}) = 1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\tilde{z}^{-m}. \tag{4}$$

Removing delay-free loops

- To remove delay-free loops from $B(\tilde{z})$, it synthesis filter is re-designed to

$$B(\tilde{z}) = 1 + \gamma \sum_{m=0}^{M} b'_\gamma(m)\Phi_m(z). \tag{5}$$

- where

$$\Phi_m(z) = \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)}, m \geq 1. \tag{6}$$

- and the filter coefficients $b'_\gamma(m)$ are obtained using a recursive formula

$$b'_\gamma(m) = \begin{cases} c_{\alpha,\gamma}(M), & m = M \\ c_{\alpha,\gamma}(m) - \alpha b'_\gamma(m+1), & 0 \leq m < M \end{cases} \tag{7}$$

A structure of MGLSA filter

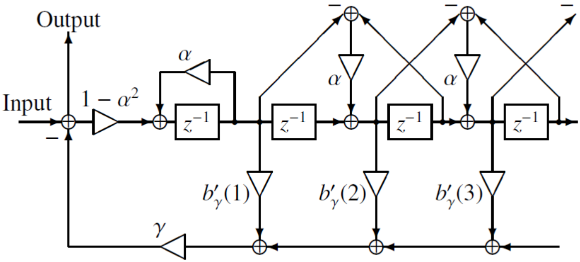

Figure: A structure of MGLSA filter $\frac{1}{B(\tilde{z})}$. (If you're not familiar with this kind of diagram, the triangles are scalers/attenuators i.e. multiply-by-constant, the plusses are adders, and the z^{-1} boxes are 1-cycle delays.)

This synthesis filter is known in literature as Mel-Generalised Log Spectral Approximation (MGLSA) filter.

Mel-generalized cepstral vocoder (MGC)

The MGC vocoder is based on analysis/re-synthesis framework introduced in the previous section. The main characteristics are:

- Uses a mixture of pulse train and white Gaussian noise for excitation source modelling.
- Pulse/noise model is straightforward.
- Produces characteristic “buzzy” sounds due to strong harmonics at higher frequencies.
- Typical parameters are $\alpha = 0.42$ and $\gamma = -1/3$.

STRAIGHT-MGC - 1999

Figure: Hideki Kawahara, Professor, Wakayama University

`http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html`

STRAIGHT: Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram

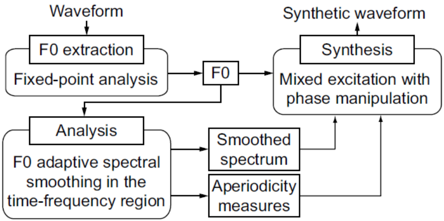

Figure: A block diagram of STRAIGHT vocoder

- Extract fundamental frequency F0
- F0-adaptive spectral analysis. The aperiodicity measure is defined as the lower envelope (spectral valleys) normalized by the upper envelope (spectral peaks).

STRAIGHT: synthesis

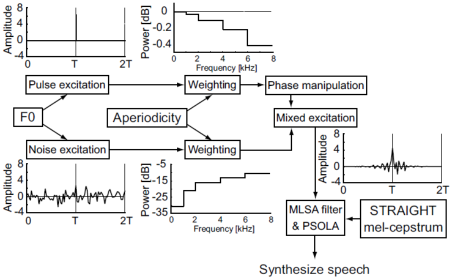

Figure: STRAIGHT synthesis

- Aperiodicity is used to weight the harmonic and noise components of the excitation; removes the periodicity effects of fundamental frequency on extracting the vocal tract spectral shape.

Glottal vocoder

- Uses a library of glottal pulses instead of pulse train for voiced signals.
- The glottal excitation is synthesized through interpolating and concatenating natural glottal flow pulses.
- The excitation signal is further modified to reproduce the time-varying changes in the natural voice source.
- Analysis of excitation using the Iterative Adaptive Inverse Filtering.
- Energy and harmonic-to-noise ratio for weighting the noise component.
- Available at `http://www.helsinki.fi/speechsciences/synthesis/glott.html`.

Deterministic plus Stochastic vocoder – 2012

- Uses MGC analysis/re-synthesis
- Differs in the excitation modelling:
  1. Uses GCI-synchronous LP residuals extraction.
  2. Deterministic component at the low frequencies is decomposed using PCA to obtain first eigen residual
  3. Stochastic component is made of energy envelope and an autoregressive model.

Harmonics plus Noise Model based vocoder - 2013

- The previous described vocoders were based on source-filter decomposition and modelling.
- An completely different approach is using sinusoidal/waveform decomposition.
- The harmonic plus noise (HNM) models assumes the speech spectrum to be composed of two frequency bands: harmonic and noise. The bands are separated by maximum voiced frequency (MVF).

HNM harmonic band analysis 1

- The harmonic part, the lower band, is modelled as a sum of harmonics

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) \exp(jk\omega_0(t)t) \qquad (8)$$

where $L(t)$ denotes the number of harmonics that depends on the fundamental frequency $w_0(t)$ and on the MVF $F_m(t)$.

HNM harmonic band analysis 2

- The complex magnitudes $A_k(t)$ can take on one of the following forms:

$$\begin{aligned} A_k(t) &= a_k(t_i) \\ A_k(t) &= a_k(t_i) + t b_k(t_i) \\ A_k(t) &= a_k(t_i) + t c_k(t_i) + t^2 d_k(t_i) \end{aligned} \tag{9}$$

  where $a_k(t_i), b_k(t_i), c_k(t_i)$ and $d_k(t_i)$ are complex numbers with constant phases, measured at analysis time instants t_i.

- A simple stationary harmonic model using the firstly defined $A_k(t)$ is referred as HNM_1 is capable to generate speech perceptually indistinguishable from original speech.

HNM noise band analysis

- The modulated noise, the upper band. Most important is specification of noise bursts (where energy is localised). Therefore the noise part $s_n(t)$ is described as time-varying autoregressive model $h(\tau, t)$ modulated by a parametric envelope $e(t)$:

$$s_n(t) = e(t)[h(\tau, t) * b(t)] \tag{10}$$

where $b(t)$ is white Gaussian noise.

- Finally, the synthetic speech $\hat{s}(t)$ is

$$\hat{s}(t) = s_h(t) + s_n(t) \tag{11}$$

HNM based vocoder

The HNM based vocoder thus:

- Decomposes the speech frames into a harmonic part and the stochastic part using
  1. MGC
  2. F0
  3. MVF
- Voiced frames – full spectral envelope may be obtained by interpolating amplitudes at harmonics.
- Unvoiced frames – analysed with fast Fourier transform.
- Available at `aholab.ehu.es/ahocoder/index.html`

Speech quality evaluation

- In the context of the last lectures about the parametric speech, i.e., the speech analysis/re-synthesis methods, one may be interested in evaluation of speech quality degradation that the methods introduce.
- We distinguish:
  1. Subjective evaluation: by asking people about evaluated stimuli. It is costly and time consuming.
  2. Objective evaluation: by using computers for that. It is cheaper, faster, but the quality depends on the test.

Classification of speech quality evaluation methods

1. Conversational quality: the quality aspects of the conversation – it is a rare test.
2. Talking quality: echo, delay and sidetone distortion.
3. Listening quality: to measure typically single quality dimension such as:
   - intelligibility
   - naturalness
   - listening effort

Subjective listening tests

- The subjective listening tests differ mainly if the reference signal is used.
  1. Non reference based tests follow *absolute category (ACR) rating* procedures.
  2. Otherwise reference based tests are called *degradation category rating (DCR)* tests.
- Both following MOS and DMOS tests are standardised by ITU-T.

Mean Option Score (MOS)

In an ACR test a group of listeners rate the listening quality of the stimuli (speech examples). The quality is rated in the 5-level impairment scale:

1. Bad,
2. Poor,
3. Fair,
4. Good,
5. Excellent.

and the average of all scores is represents the speech quality metric called mean opinion score (MOS).

Degradation Mean Option Score (DMOS)

- Sometimes the resolution of the MOS is not sufficient. It can be increased by reference based DCR test.
- Here the listeners first listen original (source) speech signal and rate the degradation of speech quality of the processed (modified) speech signal. The degradation is again rated in the 5-level impairment scale:
  1. very annoying,
  2. annoying,
  3. slightly annoying,
  4. audible but not annoying,
  5. inaudible.
- The average of all scores is represents the speech quality metric called degradation mean opinion score (DMOS).

ABX

- If one can test listener' reliability as well, there is so called ABX test.
- The listeners are provided with three speech examples – A, B, and X, asking which of A/B is identical to X. As the signal X is known reference, the ABX test also belongs to the DCR procedures.
- The ABX test is suitable for rating small degradation using a continuous impairment scale, and expert (trained) listeners should be used.

Objective

1. Similarly as in subjective listening tests, reference based tests are called *intrusive*.
2. Non reference based are called *non-intrusive*.

Spectral distortion

Widely accepted objective measure is a frequency domain measure – gain-normalised spectral distortion (SD). The SD measure evaluates autoregressive spectra $P_{xy}(n,k)$

$$P^R_{xy}(n,k) = < R_{xy}(k), \exp^{-j2\pi nk/N} > \tag{12}$$

as per frame k

$$d^k_{SD}(s,t) = \frac{1}{N} \sum_{n=0}^{N-1} \left[10 \log_{10} \left(\frac{P^s_{xy}(n,k)}{P^t_{xy}(n,k)} \right) \right]^2 \tag{13}$$

for the source signal s and target signal t. The final measure, the global distortion is the root-mean SD:

$$d_{SD}(s,t) = \frac{1}{K} \sqrt{\sum_{k=0}^{K-1} d^k_{SD}(s,t)} \tag{14}$$

where K is the total number of frames.

Psycho-acoustically motivated measures

Many of the intrusive objective measures are psycho-acoustically motivated measures. The idea here is to mimic human speech listening, and so the methods implement two basic modules:

1. Auditory processing – it employs an perceptual transform using bark-scale frequency warping and subjective loudness conversion. The output is the auditory (nerve) excitation.
2. Cognitive mapping – it extract key information related to anomalies in the speech signal from the auditory excitation. This area is still not well understood.

Perceptual Evaluation of Speech Quality (PESQ)

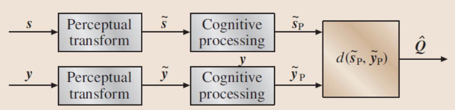

Figure: Mimicking human quality assessment.

- Widely used Perceptual Evaluation of Speech Quality (PESQ) computes internal representations based on auditory periphery of both reference/source signal s and distorted/target signal y
- Internal representations are compared to predict speech quality degradation $\hat{Q}$.
- It mimics human brain that probably compares these two entities during speech quality evaluation as well.

Perceptual Objective Listening Quality Assessment (POLQA)

A recent update of PESQ measure is Perceptual Objective Listening Quality Assessment (POLQA):

- PESQ measures one-way distortion and the effects related to two-way communication such as delay, echo are not reflected in the scores. POLQA handles the signal with variable delays.
- PESQ was design for narrow-band signal (3.4 kHz) and even there is an wide-band (7 kHz) extension, POLQA should perform better for wide-band signals

POLQA enhancements

- POLQA in addition predicts "idealised" reference signal, modelling listeners expectations of an ideal signal.
- The reference signal with low amount of recording noise and an identical degraded signal will not be scored with the maximum score.
- When the uncertainty of the subjective scores is taken into account, a statistical metric called *epsilon-insensitive rmse (rmse*)* can be used (ITU-T P.1401 (07/2012)).
- Last but not least: PESQ is free while a binary of POLQA costs 3500 CHF.