Unwrapping Codec 2 Data Manifolds

November 6, 2023

> Uncrumpling paper balls is what machine learning is all about: finding neat representations for complex, highly folded data manifolds
>
> Francois Chollet [**?**]

Speech coding can be defined as the art of representing speech using a small number of parameters that can be efficiently sent over a channel.

Codec 2 700C resamples the pitch dependent, time varying $L = F_s/(2F_0), L = 20..80$ harmonic magnitude samples $\{20log_{10}(A_m)\}, m = 1..L$ to a set of fixed length $K = 20$ samples $\{B_k\}, k = 1..K$. These are then vector quantised $\hat{B_m} = Q(B_m)$. Harmonic phase is recovered using a minimum phase model. The recovered phase spectra and hence energy distribution across the pitch cycle is dependent on the magnitude spectra available at the receiver.

There is some evidence that the distribution of energy over a pitch cycle is important for speech perception [**?**]. To adequately represent male speech, the formant energy must be spread in time. As Codec 2 recover phase using a minimum phase model - this implies narrow formant bandwidths must be maintained [1]. Narrow formant bandwidths lead to long decay times for filters "ringing" with formant energy. When low resolution spectral sampling is employed (e.g. smoothed vector quantised spectra or low bit rate LSP quantisation of LPC models), male speech becomes buzzy, muffled and less intelligible. Speech with a short pitch period (e.g. female) has less time to decay between pitch cycles so is less sensitive to formant bandwidths, the short pitch period naturally leads to a more uniform distribution of energy over time.

At the Codec 2 receiver, a post filter $P()$ is applied which improves the perceptual quality, in particular for low pitched (male) speakers. The post filter is a form of ad-hoc non-linear filtering, using experimentally derived constants. It raises formants and lowers anti-formants, effectively reducing formant

[1] Other ways of spreading time domain energy are possible, for example generating phases through heuristics, non-minimum phase models, or time domain compression [**?**]

bandwidth and more widely distributing energy in time (each formant "rings" longer).

Neural codecs have shown high quality speech can be synthesised using vectors of sparsely sampled frequency domain samples (MFCCs) [**?**]. Applying linear transformations (such as IDCT and interpolation) to recover $\{A_m\}$ from MFCCs results in poor quality synthesised speech in sinusoidal codecs. This implies neural codecs are employing a non-linear transformation, made possible by modern deep learning techniques.

From an information theory perspective there is no reason to believe there is more information in speech from low pitch speakers than high pitched speakers. Therefore with an appropriate transformation we should be able to synthesise equivalent quality speech regardless of pitch.

This document describes two experiments to determine if narrow bandwidth formants can be preserved for low pitch speakers using non-linear transformations, resulting in reasonable quality speech when synthesised using Codec 2 (without the use of a post filter).

1 Vector Quantisation

Vector quantisers can take advantage of linear and non-linear dependencies to reduce bit rate [**?**]. They can efficiently exploit information that is uncorrelated in a linear sense (no linear transformation exists), but is statistically dependant.

The 700C resampling to a fixed length $K << L_{max}$ is a linear transformation aimed at fixing the dimension of the data; reducing the distortion with multi-stage VQ where residuals tend towards independent Gaussians; and reducing VQ storage requirements. However the resampling is a filtering/smoothing/aliasing operation, so some information is lost, for example narrow bandwidth formants cannot be recovered by a linear transformation.

In this experiment we will upsample the variable rate L vectors to a fixed length $K = 80$, then vector quantise. In this way no information is lost from the source vectors. The aim is to utilise the non-linear dependency matching properties of VQ to preserve high quality speech, while still maintaining a similar bit rate to $K = 20$ quantisation. We will ignore storage concerns for this experiment.

2 Machine Learning

In this approach we will attempt to take a $K = 20$ vector and resample it to $K = 80$ using a small neural network. We hope the network will discover any non-linear dependencies, and produce narrow bandwidth formants and (when synthesised using Codec 2) reasonable quality speech for low pitched speakers.

We will include pitch as a feature (perhaps via an embedding network), arguing that the formant bandwidth is a function of pitch (F_0). This is information the VQ in Section 1 does not consider, which implies better expected results

from the ML approach. Using a fixed length target vector simplifies any issues around variable vector length. This is similar to the decoder side of an autoencoder network. Autoencoder designs could therefore be used as candidate architectures.

We argue that essentially the same information (in the form of MFCCs) is used for high quality speech synthesis with neural vocoders, therefore these networks must be performing a similar non-linear mapping of coarsely sampled, smoothed spectral information to speech spectra that includes narrow bandwidth formants. A counter argument is a simple neural network may not be capable of representing the non-linear function that maps between the two vectors. However we are not trying to synthesise speech here - just the speech spectral envelope. Other differences: the network proposed here outputs linear values using regression, more sophisticated neural vocoders employ sophisticated conditional probability models and output a PDF; this network considers just one frame, rather than utilising an autoregressive design that considers many past samples.

3 Evaluation

Evaluation methods considered:

1. Informal listening tests and ranking across a small number of samples.
2. Evaluation of the results using objective measures such as Spectral Distortion (SD), however the relationship to subjective quality may be complex. For a given bit rate SD may be larger at $K = 80$ but the synthesised speech may sound better.
3. Another candidate objective measure is Peak Average Power Ratio (PAPR). This tends to be higher when the formant bandwidths are not preserved.
4. Visual inspection of speech spectra and waveforms of speech synthesised from both methods (prior to post filtering) would indicate if narrow formants have been preserved for males.

Success would be indicated by evidence of formant bandwidth being preserved, and higher quality speech from male speakers compared to linear $K = 20$ approaches. Speech should be synthesised using phases recovered from the output magnitude spectra using the minimum phase model - the degradation in quality is less obvious when original harmonic phases are used.

For controls we could use:

1. Speech resampled through the $K = 20$ "bottleneck" (without postfiltering).
2. Speech VQed at $K = 20$ and $K = 80$.
3. Codec 2 3200 - this uses finely quantised LSPs that preserve formant shapes (and a post filter).

4 $K = 80$ VQ Results

$K = 20$ and $K = 80$ 2x12 bit (24 bit total VQs) were trained on 114,000 vectors using the LBG algorithm. A single male and female sample were processed under a variety of conditions (see `README`). The $K = 80$ male sample sounded better than the $K = 20$ sample, however there is still significant distortion. A significant artefact is a clicky/buzzy attribute after VQ, which is much more obvious at $K = 20$. The female sample was approximately the same for both $K = 20$ and $K = 80$.

Figure 1 is a plot of a single frame from the male sample. The SD was significantly higher for $K = 80$ than $K = 20$, the reverse of the subjective results. Note the narrow bandwidth of the formants - in this frame F1 is defined by a single harmonic around 10dB higher than adjacent harmonics. While $K = 80$ preserves the spectral detail well, the VQ is very noisy due to the difficulty with large dimension, multi-stage VQ. The largest errors (black) are around the formants - precisely the opposite of what we would like.

Visual inspection of the male time domain synthesised waveforms and measurement of PAPR (Figure 2) demonstrates the energy distribution issue. The process of dimension reduction and quantisation leads to spectral smoothing; narrow formants become wider. A wider filter in the frequency domain leads to more concentration of energy in time (faster decay of the time domain envelope), which we believe is the source of the unpleasant buzzy artefact. The human ear does not like clicks.

These experiments suggest the $K = 80$ VQ is better at preserving spectral detail such as narrow bandwidth formants and provides a better time domain energy distribution than $K = 20$. However large dimension multi-stage VQ with acceptable quality is difficult, as the residual error vectors from each stage tend towards independent Gaussians.

The problem remains - higher pitched speakers sound acceptable at low spectral resolution such as $K = 20$, so it should be possible to encode low pitch speech with acceptable quality at the same dimension.

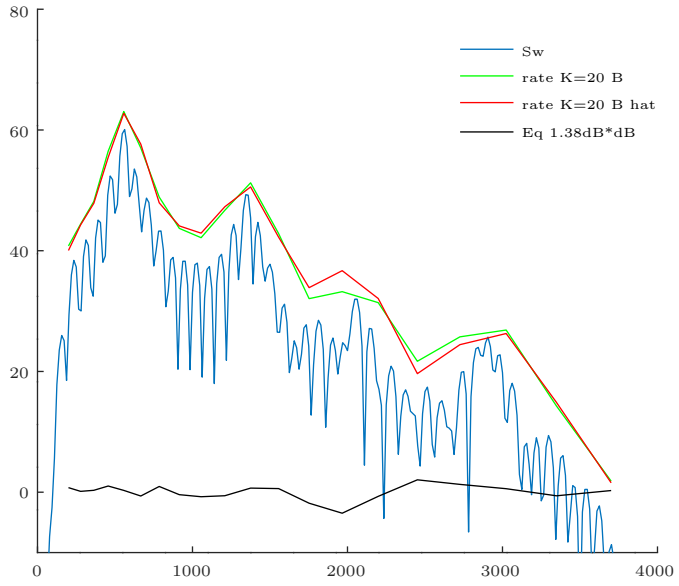
5 ML Design

In order to develop a ML system, we need to develop a cost function and appropriate model. A first choice is $K = 80$ vectors as a target, and MSE between log spectra as the cost function. However like VQ, this will try to minimise the error across the entire spectra, and not focus on the critical formant bandwidths. The VQ search uses MSE between log spectra, and often removes/rounds off spectral peaks and broadens formant bandwidths. VQ training using kmeans also encourages averaging which de-emphasises vectors with tight spectral peaks (VQ doesn't consider F_0).

Notes:

1. Matching the time domain waveform or frequency domain amplitude and phase is not critical, it's the time domain envelope that counts.

Figure 1: Voiced frame 165 frequency domain from male speaker for $K = 20$ and $K = 80$, red is 24 bits/frame VQ. Objective VQ distortion is higher at $K = 80$ however formants are sharper and formants better preserved. F3 is poorly represented in both cases. Note the formants F1-F3 are defined by single harmonics, indicating narrow bandwidth.



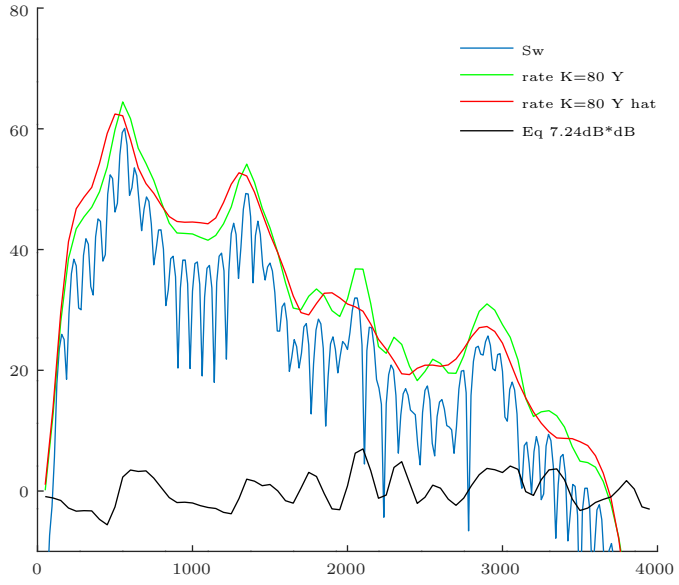
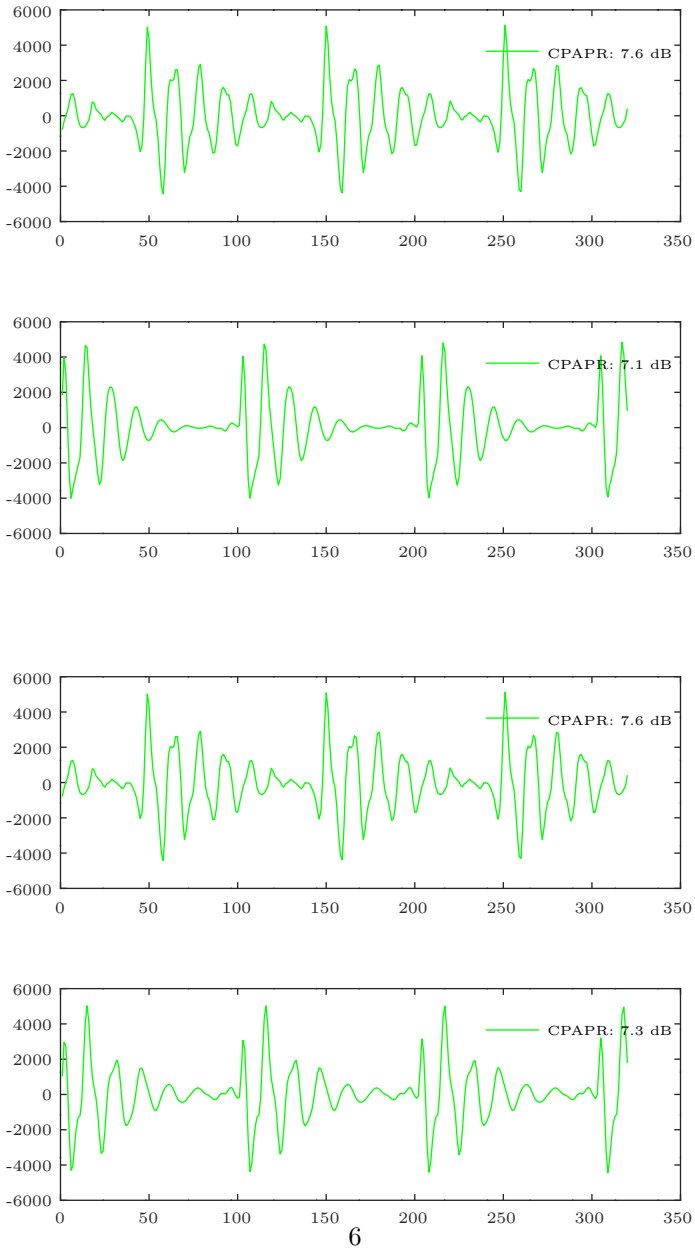


Figure 2: Voiced frame 165 time domain synthesised speech from male speaker. Plots 1 and 3 are the speech synthesised with original harmonic magnitudes and phases, presented as a control. Plot 2 is speech synthesised from the $K = 20$ model, plot 4 $K = 80$. The difference in energy distribution over the pitch cycle is clear. The objective PAPR measure ranks the samples correctly but appears only weekly related to energy distribution.



2. Lets separate the problem of quantisation from energy distribution. Use a low dimension, unquantised $K = 20$ or MFCC vector as input, and focus on reasonable quality low pitched speech as the output.

3. We are not tied to the current Codec 2 algorithm of converting magnitude spectra to phase spectra via a minimum phase Hilbert Transform. We can generate phase spectra using other means.

4. The cost function should be weighted against clicks in the synthesised speech. This implies we need a function $X(\{A_m\}, \{\theta_m\})$ that measures energy distribution.

5. PAPR is not satisfactory (confirm/explore why).

6. We would like formant bandwidth to be a function of F_0, as it matters for low pitched speakers (e.g. males) than females. Having said that - females undersample the spectra which just one harmonic per formant so formant bandwidth is perhaps irrelevant. We just wish to preserve the power of the harmonic(s) in that formant (something the current ad-hoc post filter doesn't explicitly do). A narrow bandwidth formant would fit both male and female speech.

7. Smooth post filtering across frames is important, we don't want abrupt changes in harmonic energy or discontinuous phases.

8. For UV speech mistakenly classified at V, we should also have energy distributed - we should never synthesise clicks. UV speech tends to have broad, flat spectral peaks, so if synthesised as V will have noticeable clicks.

9. When formants are broadened, multiple harmonics of similar level are synthesised, creating a "two tone" time domain waveform envelope, rather than a damped exponential (bandpass figures).

10. Idea: can we "fit" damped exponentials to spectra, like VQ-VAE fits Gaussians?

Requirements for cost function:

1. Models energy distribution of real world speech, and way ear integrates energy.

2. Good distribution in each formant, this may not be obvious in non-preemph speech due to spectral tilt.

3. Formant energy should be the same after processing.

4. Formant energy should be smooth between frames. Avoid amplitude modulation at the frame rate.

5. Smooth transition between frames, don't break phase track.

6. Deals with time varying number of harmonics, e.g. via interpolation of sparse functions

7. Stretch goal: For UV speech mistakenly classifed as V, doesn't generate clicks.

Candidates techniques:

1. Damped exponential α.

2. Match spectra, especially around formants, and use a HT min-phase to get phases. N_b filtering broadens peaks and chops off tops, several harmonics enter formant filter and α is reduced.

3. Modify phases, not amplitude spectra. Experiments show original phase and rate K spectra sounds OK. Ear is not that sensitive to frequency as long as energy distribution is OK. Modifying spectra can lead to other artefacts as energy bounces around. OTOH, it might help intelligability.

4. Use a CELP style weighting filter. This will combine energy match of highest amplitude spectra, with some match of lower energy, and will ignore low energy HP/LP sections that have been filtered out.

5. Ratio of minima/maxima of envelope. Someone must have worked out expressions for these? Can we solve anlaytically? Fourier series stuff really.

6. Can we specify mag envelope and inverse transform to get phases?

5.1 Notes and Further Work

Notes:

1. Do we need an embedding layer for pitch? Research embedding networks.

2. Should we test with some contrived data? Introduce expected non-linearity and see if network can train as a check.

3. Just include voiced frames? This seems like a good idea. For UV frames pitch may not be meaningful, and may cause random noise. Set UV frames to a constant F_0 (like 0) so at least the network "knows".

4. Make sure we use energy normalisation [**?**] when resampling to rate K.

5. Used mean removed vectors, just consider $200 < f < 3600$ Hz, for consistency with recent Codec 2 VQ work in [**?**].

6. The $K = 80$ are linear spaced, $K = 20$ Mel spaced.

Further work:

1. If fixed K works, try to come up with a network that handles sparse vectors.

2. Are MFCCs good choices for input features to neural vocoders? They are smoothed, linear transformations of the input speech. Other latent spaces may be more suitable.

3. A linear transformation (like a DCT or PCA) followed by VQ is a good idea - truncating DCT coeffs will reduce the dimension, making storage easier, and VQ will capture any non-linear dependencies. Smaller dimension will help residual error problem in multistage VQ. Trade off between truncation and VQ distortion.

4. Can we avoid gradient shift/tendancy towards Gaussians with multistage VQ? Does mbest help?

5. Scripts supporting this study need refactoring.