Harmonic tunnelling: tracking non-stationary noises during speech

Douglas Ealey, Holly Kelleher and David Pearce

Human Interface Lab, Motorola Labs
Basingstoke, UK
doug.ealey@motorola.com

Abstract

This paper presents a novel noise robust front-end algorithm and evaluates its performance on the Aurora 2 database. Most algorithms aimed at improving the performance of recognisers in background noise make an estimate of the noise spectrum that is then used to obtain an improved estimate of the spectrum of the underlying speech. In the case of stationary noises it is sufficient to take an average noise spectrum from the period before the speech utterance and/or to use a speech/non-speech detector to update this estimate using the noise sampled from any gaps in an utterance. For non-stationary noises where the noise spectrum changes faster than the duration of a typical utterance (e.g. within 0.5s) then there can be substantial differences between the estimated and actual noise spectrum for a particular frame, leading to poor performance.

The algorithm presented here provides an improved estimate of the noise that can be tracked throughout the duration of the speech utterance, by making use of the harmonic structure of the voiced speech spectrum. This running estimate of the noise is obtained by sampling the noise spectrum in the gaps (or “tunnels”) between the harmonic spectral peaks.

Compared to the ETSI standard MFCC front-end [1], the proposed algorithm delivers an average improvement in performance of 43.93% on the Aurora 2 database [2].

1. Front-end Algorithm Overview

A summary of the algorithm is shown in figure 1 and the details of the processing performed by each block described in the section 2.

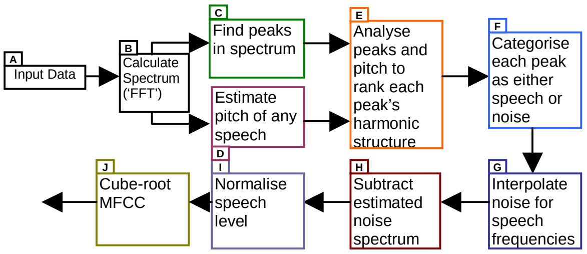

A
Input Data
B
Calculate Spectrum ('FFT')
C
Find peaks in spectrum
D
Estimate pitch of any speech
E
Analyse peaks and pitch to rank each peak's harmonic structure
F
Categorise each peak as either speech or noise
G
Interpolate noise for speech frequencies
H
Subtract estimated noise spectrum
I
Normalise speech level
J
Cube-root MFCC

Figure 1: Block diagram of front-end algorithm

The spectrum of the signal is first obtained by taking an FFT. The peaks in this spectrum are then determined from spectral derivatives. Each of these candidate peaks are analysed to categorise them as a peak coming from either a voiced speech harmonic or noise. The noise spectrum at a peak categorised as speech is estimated by interpolation from the adjacent noise spectra in the surrounding “tunnels”. These frame based noise measurements contribute to the running average of the noise spectrum in the Mel domain. Whilst this noise estimate could be used in many alternative algorithms, in this implementation it is used for an SNR dependent spectral subtraction. The remaining processing blocks are spectral normalisation performed in the Mel domain, normalising with the long term average of the spectrum and also by the frame energy. Finally a cube root compression is performed followed by cosine transform to produce 12 cepstral coefficients and a log energy measure.

2. Front-end Algorithm Details

2.1 Peak detection

The locations of the peaks in the spectrum are determined by finding the local gradient of the spectrum. A modification to this method is the use of two scales over which the gradient is evaluated, for example, 5 frequency bins and 3 frequency bins. The purpose is to discriminate in favour of significant (speech) peaks using the larger scale, and use a fractionally weighted contribution from the smaller scale differentiation to resolve the precise position of the peak. This provides low-pass filtration using the wider gradient to discriminate against random noise peaks, whilst using the finer gradient to maintain accuracy.

2.2 Pitch detection

Autocorrelation is performed on the first 40 bins of the narrowband spectrum. The peak detector of section 2.1 is then re-used to find the number of peaks and the frequency bin of the peak corresponding the highest harmonic within the autocorrelation. In this way the best fractional estimate of the pitch is obtained. Because the accuracy of the estimate is dependant on the harmonic density with these 40 bins, the cumulative pitch estimate error at the *n*th harmonic is pitch-independent, making error tolerances predictable and simplifying correction.

2.3 Analysis of the noisy speech

In figure 2 overleaf, step E of the algorithm (c.f. figure 1) is expanded.

2.3.1 Structural Analysis: Local Periodicity

In step E1, every candidate peak is given a score according to how closely its neighbouring peaks match the positions anticipated for them using the estimated pitch. A simple set of

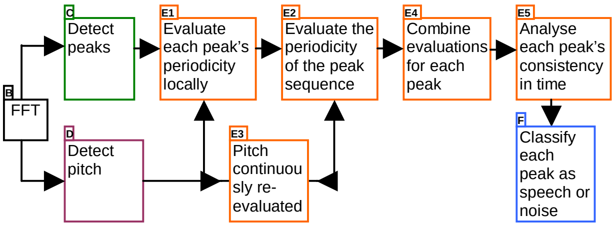
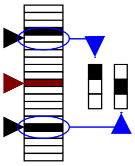
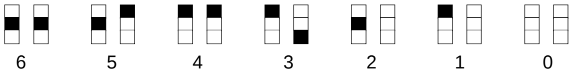

B
FFT
C
Detect peaks
D
Detect pitch
E1
Evaluate each peak's periodicity locally
E2
Evaluate the periodicity of the peak sequence
E3
Pitch continuou sly re-evaluated
E4
Combine evaluations for each peak
E5
Analyse each peak's consistency in time
F
Classify each peak as speech or noise

Figure 2. Structural analysis system diagram



rules generates the scoring criteria: in the figure opposite, the candidate peak and its closest neighbours are shown on the left. Their expected relative positions (according to the f_o estimate) are shown by arrows. The block of three frequency bins centred on these positions are then shown side-by-side on the right, forming a characteristic pattern of harmonic error. The permutations of this error type are then simply the patterns possible in both planes of symmetry. A possible score scheme is given below, with each characteristic error pattern scoring the same in both planes of symmetry:



One can see that deviation from the expected position is scored both in terms of absolute distance and consistency within the local sequence of three peaks.

2.3.2 Structural Analysis: Sequential Periodicity

As step E1 evaluates each peak locally for periodic structure, unwanted artefacts such as creak (a half-period phenomenon seen in some female talkers) may score well. The second scoring method seen in step E2 discriminates against such occurrences by scoring according to the consistency of harmonic structure over the whole spectrum. Rather than evaluating every peak individually, this method starts at the fundamental frequency and then looks for the next harmonic peak within ±1 frequency bin of its expected position. The new peak receives a score proportional to its deviation from the expected position. If the deviation is small relative to the pitch, the process continues using this new peak as the start position. Where no peak is found, the algorithm looks 2, 3, 4 etc. periods higher until a peak is encountered. The scores from steps E1 and E2 are then combined.

2.3.3 Structural Analysis: Pitch Re-Estimation

Because it is possible that part of a harmonic sequence is lost in noise, as noted in step E2 it may be necessary to predict small sequences of harmonic positions. As a consequence it is desirable that the estimate of f_o is as good as possible. The initial estimate only used peaks up to 800Hz. Consequently, when a peak at a higher frequency achieves a maximum score according to the methods described above, step E3 is used to re-evaluate the pitch period to a higher fractional accuracy.

2.3.4 Structural Analysis: Temporal Consistency

Consistency in both time and frequency requires a two-dimensional analysis of the scores derived above. This in turn requires the storage of the scores for the 'past', 'current' and 'future' frames (in effect requiring a one frame lag) to provide the context in which to evaluate the 'current' frame.

In step E5, each peak in the current frame is analysed using a mask that discriminates in favour of speech harmonic trajectories within the time-frequency space. The new score for the current peak consists of a combination of the scores of all those peaks that fall within the mask, as seen below:

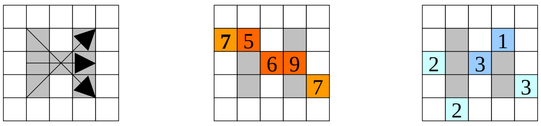


Figure 3. Left: *2D Analysis area showing likely trajectories.* Centre: *Example of a harmonic structure scoring 20.* Right: *Example of unstructured noise scoring 4.*

In figure 3, the mask is shown with two additional frames 'before' and 'after' for illustrative purposes only. One sees that in the case of speech, high scores from steps E1 and E2 fit the mask well, giving high combined scores. However, noise peaks tend to score poorly in steps E1 and E2, and then also fail to fit the mask well. Consequently combined noise scores tend to be proportionately much lower than those for speech, improving the separation of the two.

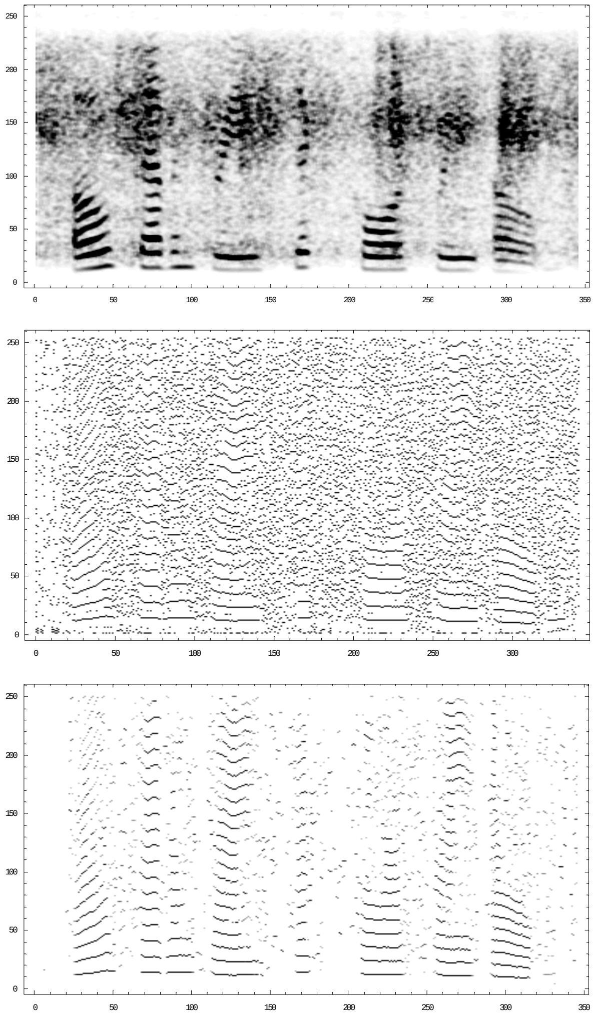


Figure 4. Top: *The utterance 'Oh-7-3-6-4-3-Oh' in 0dB SNR train noise.* Middle: *Identification of all energy peaks with no discrimination applied.* Bottom: *Remaining candidate speech peaks following application of the processes described in this paper.*

2.3.5 Structural Analysis: Classification

A threshold is selected that best differentiates between typical speech and noise scores. A modification of this method is to use two thresholds in conjunction with a whole-frame speech/non-speech detector. In this form, a high threshold is used during frames judged to be noise, and a lower one used during frames judged to be speech. In each case if a peak exceeds the applied threshold, it is deemed to be speech.

The overall effect of the peak detection and structural analyses can be seen in figure 4 above.

2.4 Harmonic tunnelling

One can see in figure 4 that the train noise is highly non-stationary, frequently changing profile during speech. A noise estimation method that relies only on the gaps between utterances will make significant estimation errors in such conditions.

To address this problem, *harmonic tunnelling* is now introduced. Given the knowledge derived in section 2.3 identifying where the speech energy is concentrated within the spectrum, it should be possible to measure the interstitial noise between these positions, thus enabling continuous estimation of the noise.

2.4.1 Identifying tunnels

The identification of harmonic tunnels (Step G) is summarised in the flow diagram below:

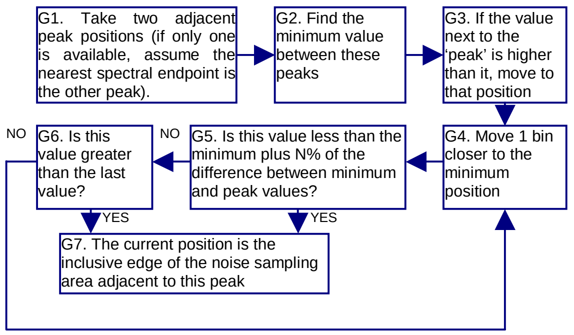

G1. Take two adjacent peak positions (if only one is available, assume the nearest spectral endpoint is the other peak).
G2. Find the minimum value between these peaks
G3. If the value next to the 'peak' is higher than it, move to that position
G4. Move 1 bin closer to the minimum position
G5. Is this value less than the minimum plus N% of the difference between minimum and peak values?
NO
G6. Is this value greater than the last value?
NO
YES
YES
G7. The current position is the inclusive edge of the noise sampling area adjacent to this peak

Figure 5. Flow diagram for tunnel identification

Step G3 in the flow diagram accounts for the existence of two-bin peaks that occur due to the frequency resolution.

Step G5 allows the setting of a 'noise floor' below which it is safe to assume all values are representative of the noise. This value can be empirical or derived using the peak value and windowing characteristics.

Step G6 detects any noise that breaks the expected monotonic profile of the peak above the noise floor level.

This process is repeated over the whole spectrum, identifying the positions of the harmonic tunnels in that frame. The values found in these tunnels are then used to interpolate the noise level for each harmonic peak, by averaging N<=4 values on each side of each peak to generate the end-points of the interpolation.

The final output is an estimate of the noise spectrum, giving exact values within the tunnels and interpolated estimates within the harmonic speech peaks, as shown in figure 6.

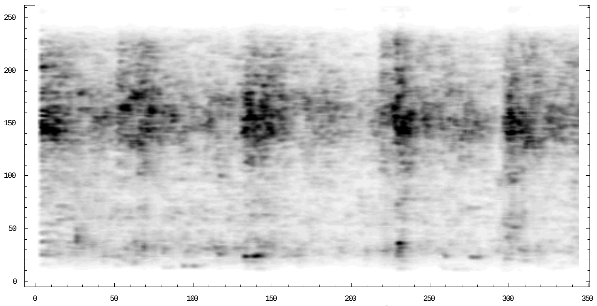


Figure 6. Harmonic tunnelling estimate of the non-stationary train noise from utterance 'Oh-7-3-6-4-3-Oh', as seen in figure 4.

3. Implementation

The data is presented in 30ms frames, advancing 10ms per frame. Pre-processing involves only the removal of the DC offset. The data is then hamming windowed and presented to a 512-point FFT. The resulting narrowband spectrum is analysed up to step E4 in figure 2 (combining harmonic scores) and then placed in a circular buffer of three frames.

Once this buffer is fully occupied (by frame 3), the central frame is analysed from step E5 (temporal consistency) onwards to generate an estimate of the narrowband noise spectrum in that frame.

This noise estimate is added to a short rolling average (where the current input contributes 25%). The smoothed estimate is then used to perform an SNR dependant spectral subtraction of the original narrowband spectrum.

The spectrum is then re-scaled to normalise the sum of the speech components within the spectrum, using the information from the harmonic tunnelling. By normalising the spectral amplitude, a non-linear function other than a log can be applied without distortion, improving the dynamic range of the final output.

The narrowband spectrum is presented to 23 Mel-filter banks, as defined in [1]. Next, a spectral normalisation based on the mean speech profile is applied. The mean profile is based on a rolling spectral average of the speech that is stored between utterance files. It is copied and updated during the current utterance and at the end contributes 1% to the stored spectral average. This enables a slowly evolving corpus level description of the speech spectrum to be built, which is particularly effective in normalising microphone characteristics.

Finally, the non-linear transformation takes the form of a cube root rather than a log, before the discrete cosine transform is performed.

The final output is 13 cepstral coefficients (of which C_0 is not used in the recogniser) plus log energy. The total latency of the system is 20ms.

Motorola UK Research Labs

Aurora 2 Multicondition Training - Results

	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage Improvement
Clean	98.80	98.58	98.63	98.77	98.70	98.80	98.58	98.63	98.77	98.70	98.65	98.46	98.56	98.67	9.47%
20 dB	98.31	98.58	98.15	98.12	98.29	97.79	98.19	97.88	97.99	97.96	98.40	97.82	98.11	98.12	28.24%
15 dB	97.73	97.40	97.85	97.59	97.64	96.99	97.34	96.63	97.28	97.06	97.70	97.64	97.67	97.42	28.59%
10 dB	95.39	95.77	96.81	95.65	95.91	94.38	95.68	94.42	95.19	94.92	95.43	95.59	95.51	95.43	25.43%
5 dB	92.08	90.24	92.22	90.03	91.14	87.47	91.20	89.74	90.34	89.69	90.27	89.12	89.70	90.27	32.13%
0 dB	78.05	66.60	71.88	74.58	72.78	67.76	73.28	72.20	69.79	70.76	70.80	67.50	69.15	71.24	29.06%
-5dB	37.70	28.81	25.53	33.97	31.50	32.70	34.89	33.16	28.57	32.33	28.03	29.08	28.56	31.24	8.90%
Average	92.31	89.72	91.38	91.19	91.15	88.88	91.14	90.17	90.12	90.08	90.52	89.53	90.03	90.50	
	31.63%	14.67%	36.05%	26.43%	27.38%	23.86%	31.63%	20.49%	34.08%	27.72%	43.42%	33.30%	38.53%		30.18%

Aurora 2 Clean Training - Results

	A					B					C				
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Overall	Percentage Improvement
Clean	99.20	99.06	99.02	99.20	99.12	99.20	99.06	99.02	99.20	99.12	99.26	99.03	99.15	99.13	9.38%
20 dB	98.53	97.34	98.27	98.12	98.07	96.71	98.28	97.55	97.99	97.63	98.04	97.94	97.99	97.88	59.45%
15 dB	96.75	94.71	97.32	96.30	96.27	93.92	96.92	96.00	96.61	95.86	96.78	96.40	96.59	96.17	70.93%
10 dB	93.46	87.61	93.77	91.58	91.61	86.40	92.68	90.61	92.22	90.48	92.02	91.20	91.61	91.16	73.40%
5 dB	83.36	70.68	81.66	80.68	79.10	69.17	81.80	75.90	79.02	76.47	80.29	78.23	79.26	78.08	64.10%
0 dB	60.49	43.17	54.13	55.94	53.43	44.03	56.50	53.59	51.84	51.49	51.43	51.21	51.32	52.23	42.37%
-5dB	28.74	18.26	20.88	24.59	23.12	17.38	26.18	22.61	22.62	22.20	21.37	22.31	21.84	22.49	15.25%
Average	86.52	78.70	85.03	84.52	83.69	78.05	85.24	82.73	83.54	82.39	83.71	83.00	83.35	83.10	
	55.82%	57.50%	62.00%	55.28%	57.82%	53.69%	61.63%	63.06%	62.89%	60.20%	51.86%	49.81%	50.84%		57.69%

Table 1. Results for the Root-Normalised Harmonic Tunnelling algorithm, giving an overall improvement of 43.93%.

4. Results

The experimental databases, evaluation conditions and metrics are described fully in [2]. The evaluation conditions for the Aurora 2 database are divided into multi-condition training and clean training, covering eight noise environments and two types of convolutional distortion.

The speech recognition experiments were done using HTK for both delta and acceleration calculations, keeping the values used in the previous standard [1]. Training and testing were performed with ETSI provided scripts. After generating 13 delta and 13 acceleration coefficients, there are 39 coefficients in total as input to the speech recogniser.

Table 1 details the results for the root-normalised harmonic tunnelling algorithm. As seen in the summary of table 2 below, the overall performance is fairly consistent at around 43% in all three test sets.

Motorola UK Research Labs

Absolute performance				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.15	90.08	90.03	90.50
Clean Only	83.69	82.39	83.35	83.10
Average	87.42	86.23	86.69	86.80

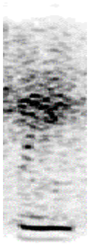
Performance relative to Mel-cepstrum				
Training Mode	Set A	Set B	Set C	Overall
Multicondition	27.38%	27.72%	38.53%	30.18%
Clean Only	57.82%	60.20%	50.84%	57.69%
Average	42.60%	43.96%	44.68%	43.93%

Table 2. Overall absolute and relative performance summaries.

5. Conclusions

In this paper we have presented an algorithm for a noise robust distributed speech recognition front-end. The algorithm is based on the novel innovation of harmonic tunnelling, which provides the continuous noise spectrum estimation necessary for SNR-dependant speech enhancement techniques such as root-normalised MFCCs.

We believe that harmonic tunnelling presents significant benefits over other methods that identify harmonic components. The most significant is that no assumption is made about the predictability of the harmonic speech SNR within noisy spectra; In the utterance of 'three' opposite, one can appreciate that a pitch-based harmonic predictor or a comb filter would primarily be sampling noise within the spectrum, since the harmonic structure *that exceeds the noise floor* is not simply all those points at multiples of the fundamental pitch. The result would be a heavily distorted estimate of the speech energy and noise profile.



By contrast, the harmonic tunnelling algorithm makes no assumptions about the presence of harmonics, and instead evaluates all structure within the signal for harmonic qualities. As a result it avoids the significant problem of obtaining *a priori* knowledge of which multiples of the fundamental to sample within noise.

The overall performance of the algorithm demonstrates that this proposal does enhance the robustness of the front-end against interfering noise and acoustic channel variation. We anticipate that the addition of voice activity detection will improve these results further by removing extraneous noise frames.

6. References

[1] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", ETSI ES 201 108 v1.1.2 (2000-02), Feb. 2000.

[2] H G Hirsch & D Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"; Paris, France, September 18-20, 2000