Improved ASR in noise using harmonic decomposition

David M. Moreno*, **Philip J.B. Jackson**†, **Javier Hernando*** and **Martin J. Russell**‡

* TALP Research Centre, Universitat Politècnica de Catalunya, Barcelona, Spain.

† CVSSP, Electronic Engineering, University of Surrey, Guildford, UK. [p.jackson@surrey.ac.uk]

‡ Electronic Electrical & Computer Engineering, University of Birmingham, Birmingham, UK.

ABSTRACT

Application of the pitch-scaled harmonic filter (PSHF) to automatic speech recognition in noise was investigated using the Aurora 2.0 database. The PSHF decomposed the original speech into periodic and aperiodic streams. Digit-recognition tests with the extended features compared the noise robustness of various parameterisations against standard 39 MFCCs. Separately, each stream reduced word accuracy by less than 1 % absolute; together, the combined streams gave substantial increases under noisy conditions. Applying PCA to concatenated features proved better than to separate streams, and to static coefficients better than after calculation of deltas. With multi-condition training, accuracy improved by 7.8 % at 5 dB SNR, thus providing resilience from corruption by noise.

1 INTRODUCTION

In a conventional front end for automatic speech recognition (ASR), incoming speech signals are converted into Mel-frequency cepstral coefficients (MFCCs), before any analysis or interpretation is carried out (e.g., by Viterbi decoding). In the present study, we have sought first to separate the voiced and unvoiced contributions to the speech signal (as periodic and aperiodic components respectively), which are then converted into MFCCs. Thus, the acoustic models may be considered as learning distinct characteristics of the voiced and unvoiced parts for any given phoneme.

The acoustic cues of speech come from a variety of different mechanisms, such as phonation, frication and plosion. Many ASR front ends treat them equally although human speech production and speech coding studies have shown the characteristics of radiated speech signals to depend greatly on vibration of the vocal cords. Standard front ends try to extract features that are not strongly influenced by the source characteristics. Here, we attempt to segregate harmonic and noise-like cues before describing their characteristics, by extracting the contribution from voicing (with large relative amplitude) from those of other acoustic sources, hence improving the feature extraction for both kinds of cue: voiced and unvoiced.

Although researchers have experimented with a plethora of ways to extract a single set of features from speech, methods of decomposing the acoustic signal from the speaker into parallel streams of information are not so well studied. Some have shown benefit in sub-band processing [1] and MFCCs mixed with formants [2], while others have used a single set of features with parallel models [3]. Since we know from personal experience that whispered, breathy or creaky speech is more difficult to understand in a noisy environment than normally-phonated speech, it seems logical that a feature extraction technique for ASR that also exploits the signal's harmonicity should offer gains in recognition accuracy and robustness to noise.

Therefore, to separate the quasi-periodic voiced component from the noise-like residual, the pitch-scaled harmonic filter (PSHF) was used. It was designed to split an input speech signal into two synchronous streams: periodic and aperiodic, which act respectively as estimates of the voiced and unvoiced components of the signal at any time [4]. After decomposition, features extracted from each of the streams may be concatenated or further manipulated into an extended feature vector, as required. The feature extraction processes are described below, with experimental details, and a brief discussion of the results, which demonstrate the capability of the PSHF for enhancing the digit-recognition accuracy of an ASR system in tests on the Aurora 2.0 database.

2 METHOD

Preparation of the acoustic features from Aurora had three main stages: (i) estimating the fundamental frequency for voiced sections of the speech corpus, (ii) decomposing the speech files into periodic and aperiodic components, and (iii) calculating the feature vectors. All training and test utterances are processed alike, as in figure 1.

ISBN 1-876346-48-5

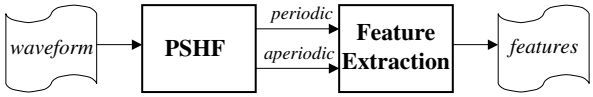


Figure 1: Front-end overview. The waveform is split by the PSHF into periodic and aperiodic components, from which features are extracted.

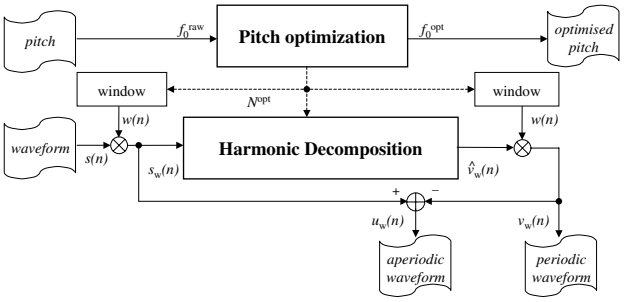


Figure 2: The PSHF (from top): optimal pitch and period, f_0^{opt} and N^{opt}, are calculated, harmonic decomposition estimates the periodic contribution, which is subtracted from the original signal to give the aperiodic estimate.

2.1 Pitch extraction

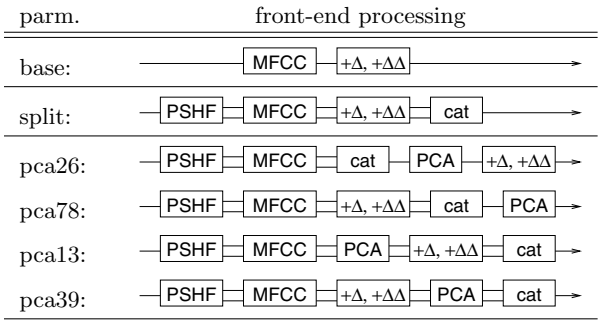
An initial estimate of each file's fundamental frequency f_0^{raw} was made, then optimised by the PSHF, which scales the window size to the pitch period as part of the decomposition. After robust pitch extraction by the Entropic utility `get_f0`, our own pitch-correction script was applied to resolve glitches in voice activity and pitch discontinuities, e.g., octave errors. The parameters of both steps were determined empirically (minimum voiced/unvoiced segment durations of 30 ms/10 ms, respectively). The clean files were processed automatically to produce f_0^{raw} values for the entire database, which the PSHF optimised with a matched cost function to yield f_0^{opt} (4 periods, 8 harmonics, 4 ms shift, [5]).

2.2 Periodic-aperiodic decomposition

The harmonic decomposition was performed using the optimised clean pitch estimates, giving a pair of periodic and aperiodic files for every file in the database. Figure 2 shows the windowing and decomposition of a frame of speech, which is by selection of harmonics in the frequency domain. Successive shift and respilicing of the outputs yields complete periodic and aperiodic signal, synchronised with the input. The algorithm and an assessment of its performance are described elsewhere [4, 6]; software and examples are online, [7].

2.3 Feature extraction

Standard MFCC features (0–12, plus deltas and delta-deltas) were extracted from the original signal and from the pair of decomposed signals, using HTK [8]. A small amount of Gaussian white noise, or dither, was added to the periodic features during voiceless sections.[1] As well as concatenation, the technique of principal component analysis (PCA) was employed to offer six parameterisations of the data,



where "$+\Delta, +\Delta\Delta$" denotes calculation of 1st- and 2nd-order differences (aka. velocities and accelerations), and "cat" implies concatenation of the periodic and aperiodic feature streams. PCA parameterisations are distinguished by the size of matrix in the analysis, which depends on the operations' order. Thus all feature vectors had 78 coefficients, except BASE with 39.

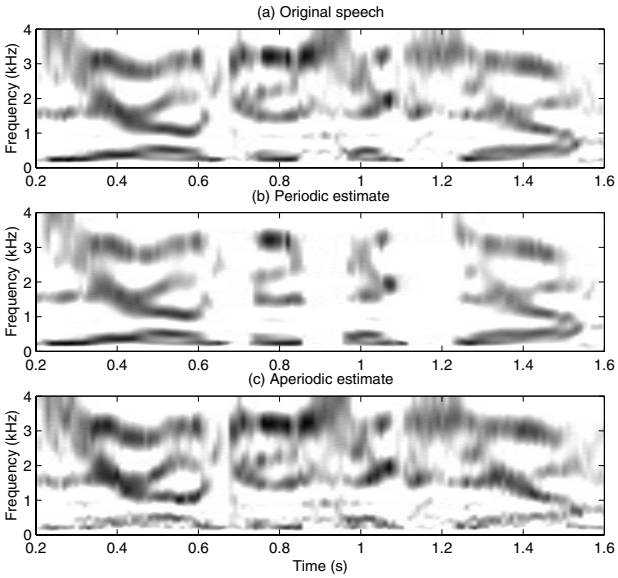


Figure 3: MFCC-derived spectrograms of the utterance "zero-two-six-zero": (a) s, (b) $\hat{v}$, and (c) $\hat{u}$.

2.4 Recognition experiments

The Aurora 2.0 database comprises clean 8 kHz speech recordings of connected digits with noise added at seven signal-to-noise ratios (SNRs): ∞, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. There are matched and unmatched noise conditions in the test data for both additive and convolutional noise (i.e., channel distortion). Hence, a recognizer may be trained using only clean data or multiple SNR conditions, and the results viewed according to test SNR.

[1] Adding dither avoided numerical instability in training probability distributions that may be induced by total silence.

ISBN 1-876346-48-5

Training scripts instructed HTK to generate a set of 16-state word models for each of the digit prototypes (and a 3-state silence model). After flat initialisation and 16 iterations of the Baum-Welch algorithm, the models were tested and word accuracy recorded. In the SPLIT tests, likelihoods of the two streams could be weighted independently. For all results reported here, the same weighting was used during training, thanks to minor adjustment of HTK [9].

3 RESULTS

Figure 3 gives a spectrographic example of the features used in the recognizer, showing the effect of the standard front end on the original signal and on the periodic and aperiodic components. Although no new information was introduced by the decomposition, it is interesting to observe the prominence of voicing transitions and the distribution of spectral details during voiced segments. From listening, the aperiodic estimate sounds similar to whispered speech (as expected for an absent voice source); though the periodic estimate contains only voiced segments, it is perfectly recognizable, due to language and coarticulation cues that remain. Under noisy conditions, the incoherent aperiodic contribution accrues most distortion and is much more easily masked than the periodic one.

3.1 Effect of decomposition

Points for equally-weighted streams, $\gamma_p = \gamma_a = 1.0$ (centre of each graph in figure 4), correspond to direct concatenation of the periodic and aperiodic features. The improvement in recognition word accuracy is remarkable, especially under noisy test conditions, suggesting that useful information had been masked in the features extracted from the original speech.

3.2 Influence of stream weights

Changing the balance of the streams weights defines three scenarios: (i) under clean test conditions, best performance was achieved when the aperiodic stream carried much more weight than the periodic one; (ii) in very noisy conditions, the best results arose with all the weight given to the periodic stream; (iii) at intermediate noise levels, a combination of both streams gave best results. This behaviour is due to the PSHF mainly ascribing corrupting noise to the aperiodic component.

3.3 Principal component analysis

PCA was used to decorrelate the dimensions of feature data, and sort them by the proportion of variance each dimension explains. It enabled us to determine which part of the variation in the data was useful to the recognizer. How complementary, or redundant, the periodic and aperiodic streams were can be measured by the number of dimensions, or PCs, beneficial to the recognition task.

Typically there were only 13 dominant dimensions in the data (including the deltas and delta-deltas) but the detection of voiced segments introduced one extra for the periodic component, so we would expect it to contain one more useful PC. With a threshold at 1 % of the total variance, the numbers of selected PCs for original, periodic and aperiodic streams were 13, 10 and 13, and 15 for the recombined streams (SPLIT). If periodic and aperiodic streams were completely redundant, the number of PCs after recombination would be equal to those for the original stream (viz. 13); if totally independent, the number should be their sum (i.e., 23). As the number of recombined PCs fell between 13 and 23, it implies that complementary information was extracted through the decomposition.

4 CONCLUSION

The PSHF was used to split each speech waveform in the Aurora 2.0 database into two synchronous streams, periodic and aperiodic, acting respectively as estimates of the voiced and unvoiced components. Features were extracted from each stream and combined (by some sequence of concatenation, PCA and calculation of delta coefficients) to form an extended feature vector. Experiments yielded accuracy scores for connected-digit recognition, and tested the noise robustness of our parameterisations against a conventional one (39 MFCCs+Δ,+$\Delta\Delta$). Used separately, each of the streams gave recognition accuracy that was only slightly degraded (by less than 1 % absolute, compared to the baseline using the original speech); together, accuracy was increased under noisy conditions, demonstrating not only redundancy between streams but also complementary information. Tests applying PCA to the concatenated feature set tended to perform better than applying PCA to the streams separately, and PCA of the static coefficients, before calculation of the deltas, was better than afterwards. With multi-condition training, the accuracy improved by 7.8 % under 5 dB SNR using concatenated streams (78 MFCCs); whereas with PCA of the combined static MFCCs, and derivatives (48 coefficients), the improvement was 5.6 %. Thus, voiced regions of a speech utterance appear to provide resilience of a message to corruption by noise. However, no significant improvement on 99.0 % baseline accuracy was achieved under clean test conditions. Complete details of this research to-date are reported in Moreno's thesis [9].

In the future, we propose to explore the influence of the voicing information on different classes of speech sound, for instance on a phoneme recognition task using TIMIT corpus, whose 16 kHz speech provides more turbulence-noise information. It would also be interesting to apply different forms of front-end processing to the two streams, and to consider other forms of model combination.

ISBN 1-876346-48-5

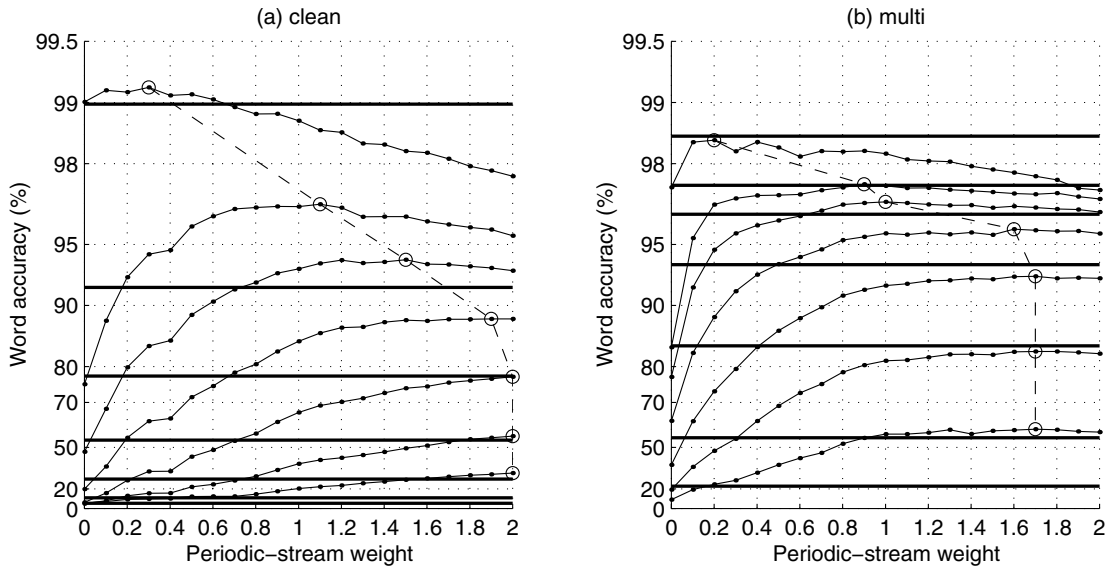


Figure 4: SPLIT test results of word accuracy (%) versus periodic-stream weight γ_p, averaged across each noise level: (a) clean and (b) multi-condition training. Solid lines are (from top): ∞, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB SNR test conditions. Thick horizontal lines indicate baseline average scores, and the dashed line (with $\odot$) indicates the best at each noise level.

	CLEAN	Signal-to-Noise Ratio (dB)							MULTI	Signal-to-Noise Ratio (dB)							
	∞	20	15	10	5	0	-5	Ave.	∞	20	15	10	5	0	-5	Ave.
base	99.0	91.9	77.7	54.0	28.4	11.4	5.8	52.6	98.5	97.4	96.5	93.7	84.2	55.2	22.4	78.3
split	99.2	96.8	94.1	88.4	77.6	56.0	33.1	77.9	98.5	97.5	96.9	95.8	92.8	83.2	59.4	89.1
pca26	99.0	95.8	92.0	82.6	64.2	40.8	23.8	71.2	98.4	97.7	97.1	95.7	92.1	81.7	59.2	88.8
pca78	98.9	94.2	87.5	70.9	44.4	23.2	14.3	61.9	98.3	97.4	96.6	95.1	91.0	80.4	57.8	88.1
pca13	98.5	96.5	93.3	85.5	68.0	43.3	23.0	72.6	98.0	97.0	96.3	94.4	90.5	79.4	57.7	87.6
pca39	98.4	95.9	91.9	83.1	64.3	39.7	23.2	70.9	97.8	97.0	96.3	94.8	90.9	79.3	56.5	87.5

Table 1: Best word accuracy (%) achieved by the each front end in § 2.3. The SPLIT and PCA results respectively depend on the stream weights and number of principal components used.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the support of their respective organizations, and Climent Nadeu, Jaume Padrell, Nick Wilkinson, Matt Stuttle and Dušan Macho for helpful discussions.

REFERENCES

[1] H. Boulard and S. Dupont, "Sub-band based speech recognition," in *Proc. IEEE-ICASSP,* Munich, 1997, pp. 1251–1254.

[2] N. Wilkinson and M. J. Russell, "Improved phone recognition on TIMIT using formant frequency data and confidence measures," in *Proc. Int. Conf. on Spoken Lang.,* Denver, CO, 2002, pp. 2121–2124.

[3] M. J. F. Gales and S. J. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Comp. Speech & Lang.*, vol. 9, pp. 289–308, 1995.

[4] P. J. B. Jackson, *Characterisation of plosive, fricative and aspiration components in speech production*, Ph.D. thesis, Dept. Electronics & Comp. Sci., Univ. of Southampton, UK, 2000.

[5] H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, and H. Fukuda, "A pitch-synchronous analysis of hoarseness in running speech," *J. Acoust. Soc. Am.*, vol. 84, no. 4, pp. 1292–1301, 1988.

[6] P. J. B. Jackson and C. H. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech," *IEEE Trans. on Spch. & Aud. Proc.*, vol. 9, no. 7, pp. 713–726, 2001.

[7] P. J. B. Jackson, D. M. Moreno, J. Hernando, and M. J. Russell, *Columbo project*, CVSSP, Univ. of Surrey, Guildford, UK, 2001, [http://www.ee.surrey.ac.uk/Personal/P.Jackson/Columbo/].

[8] S. J. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Entropic Camb. Res. Lab., Cambridge, UK, v2.1 edition, 1997.

[9] D. M. Moreno, "Harmonic decomposition applied to automatic speech recognition," M.S. thesis, Universitat Politècnica de Catalunya, Barcelona, 2002.