Covariation and weighting of harmonically decomposed streams for ASR

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

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

[†] TALP Research Centre, Universitat Politècnica de Catalunya, Barcelona, Spain.

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

Abstract

Decomposition of speech signals into simultaneous streams of periodic and aperiodic information has been successfully applied to speech analysis, enhancement, modification and recently recognition. This paper examines the effect of different weightings of the two streams in a conventional HMM system in digit recognition tests on the Aurora 2.0 database. Comparison of the results from using matched weights during training showed a small improvement of approximately 10 % relative to unmatched ones, under clean test conditions. Principal component analysis of the covariation amongst the periodic and aperiodic features indicated that only 45 (51) of the 78 coefficients were required to account for 99 % of the variance, for clean (multi-condition) training, which yielded an 18.4 % (10.3 %) absolute increase in accuracy with respect to the baseline. These findings provide further evidence of the potential for harmonically-decomposed streams to improve performance and substantially to enhance recognition accuracy in noise.

1. Introduction

The speech signal that emanates from a human speaker is typically the filtered combination of a number of acoustic sources: from voicing, frication, aspiration, and plosive release. Often more than one of these occurs at the same time, and although they tend to be quite different in nature, can be difficult to distinguish afterwards and to separate artificially. Their combined effect is what produces the full sensation of natural speech that has been a challenge for model-based speech synthesis systems. In contrast, speech recognition systems aim to take advantage of whatever useful information exists within the acoustic signal for the sake of increasing accuracy and robustness against interference. However, if we treat all sounds as one, and generate our features from parameterising sections of the speech signal based on the short-term magnitude spectrum, as per Mel-frequency cepstral coefficients (MFCCs), then we are essentially neglecting all information about the source characteristics, bar the overall amplitude and broad spectral envelope. While delta (aka. difference) features can help to identify transient cues, as with plosion or voice onset, the harmonicity of voiced speech has a quality or "acoustic texture" that sets it apart from other sounds. This paper describes the preliminary findings of an attempt to exploit the quasi-periodic property of voiced speech to improve speech recognition accuracy in a connected digit task.

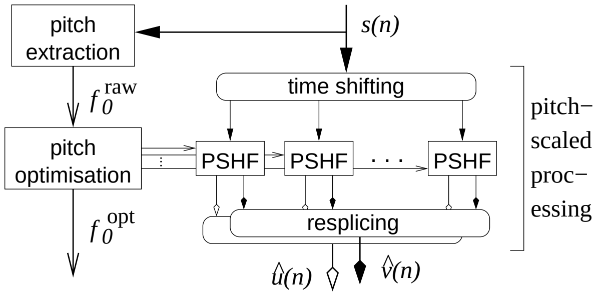

Figure 1: PSHF overview. Using the optimised pitch estimates f_0^{opt}, the waveform $s(n)$ is split by the PSHF into periodic and aperiodic components, $\hat{v}(n)$ and $\hat{u}(n)$ respectively, from which features are subsequently extracted.

The decomposition of speech signals into simultaneous periodic and aperiodic components has been successfully employed for a variety of purposes in the past. In the fields of speech modification and enhancement, examples include work by Laroche et al. [1], and by Yoo and Lim [2], respectively. Yet, the technique used in the present study was originally envisaged for acoustic analysis of speech [3], which the authors have developed to be able practically to process an entire speech corpus. Here, we report work involving the Aurora 2.0 database of clean and noise-corrupted spoken digits.

In a conventional front end for automatic speech recognition (ASR), incoming speech signals are converted into MFCCs, before any analysis or interpretation is performed (e.g., by Viterbi decoding), which has defined the baseline reference for the present study. As depicted in figure 1, we have sought first to separate the voiced and unvoiced contributions to the speech signal (as periodic and aperiodic components, $\hat{v}(n)$ and $\hat{u}(n)$, respectively), which were then converted into MFCCs. Thus, the acoustic models may be considered as learning the separate characteristics of the voiced and unvoiced parts for any given phoneme. Although many ways of extracting a single set of features from speech have been investigated, methods of decomposing the acoustic signal into parallel streams of information are not so well studied. Some have shown gains from sub-band processing [4] and mixing MFCCs with formants [5], while others have used a single set of features with parallel models [6]. From one's personal experience is it plain that creaky or whispered speech is more difficult to understand in a noisy environment than normally-phonated speech, which suggests that exploiting the signal's harmonicity should offer benefits in terms of recognition accuracy and robustness to noise for ASR.

Thus, the technique used to separate the quasi-periodic

Thanks are due to Climent Nadeu, Jaume Padrell, Matt Stuttle, Nick Wilkinson, Dušan Macho and Steve Young for helpful discussions. The Columbo project is supported by participating institutions. See http://www.ee.surrey.ac.uk/Personal/P.Jackson/Columbo/ for details.

voiced component from the noise-like residual was the pitch-scaled harmonic filter (PSHF), which was designed to split an input speech signal into two synchronous streams: periodic and aperiodic, acting respectively as estimates of the voiced and unvoiced components [3]. After decomposition, features extracted from each of the streams can be concatenated or further manipulated as an extended feature vector.

Since the vocal tract will have been in the same configuration at the instant that the voiced and unvoiced contributions were produced, there will be similarities in their respective filter characteristics. So, at least the poles of the vocal-tract transfer function, that correspond to the formants and vocal-tract resonances, would be common. Along with other factors, this implies that there would be a strong correlation between the features of the two parallel streams. The present study explores principal component analysis (PCA) as a means of removing the correlation and ensuring the efficacy of the diagonal-covariance Gaussian mixture models to represent the probability distributions of the data. More sophisticated transformation schemes, such as linear discriminant analysis or heteroscedastic LDA [7] were not considered in the work reported here. Nevertheless, a number of different schemes for calculating the transformation matrix were, and their analysis of the variance examined. An issue critical to the value of the decomposed periodic and aperiodic streams concerns the extent to which the covariation brings new information to the recognition process, meanwhile isolating the influences of disturbance from unwanted noise.

The pitch and feature extraction processes are described below, with experimental details, and a brief discussion of the results, which include a study of the effect of adjusting the weights of the two streams in tests, and of using matched or unmatched weights during training. A summary of the PCA study is also given, before concluding.

2. Method

2.1. Pitch and feature extraction

Preparation of the training and test data consisted of three steps: (i) estimate the fundamental frequency for voiced sections of the speech corpus, (ii) decompose the speech files into periodic and aperiodic components, and (iii) calculate the acoustic feature vectors. An initial robust estimate of the fundamental frequency f_0^{raw} was made by the Entropic utility `get_f0`, corrected, and then optimised by the PSHF's own cost function to give f_0^{opt} (4 periods, 8 harmonics, 4 ms shift). The pitch-correction program resolved glitches in voice activity and pitch discontinuties, e.g., octave errors (i.e., $\times\frac{1}{2}$, or $\times 2$). The parameters of both steps were determined empirically (minimum voiced/unvoiced durations of 30 ms/10 ms). The clean speech files provided f_0^{raw} values for the entire database.

The harmonic decomposition was performed from the optimised clean pitch estimates, giving a pair of periodic and aperiodic files for every file in the database. The algorithm and an evaluation of its performance are described elsewhere [3, 8].

Standard 39-dimensional MFCC feature vectors (0th to 12th, plus deltas and delta-deltas) were extracted from the original signal and from the pair of decomposed signals, using HTK [9]. A small amount of Gaussian white noise, or dither, was added to the periodic features during voiceless sections.[1] Figure 2 illustrates the features used in the recognizer spectrographically, showing the effect of the standard front end on the original signal and on the periodic and aperiodic components. It is interesting to see how decomposition highlights the voicing transitions and distribution of spectral details during voiced segments. (Further examples are on the project website.) As well as simple concatenation, PCA was used to make a total of six different parameterisations of the data, as described in table 1.

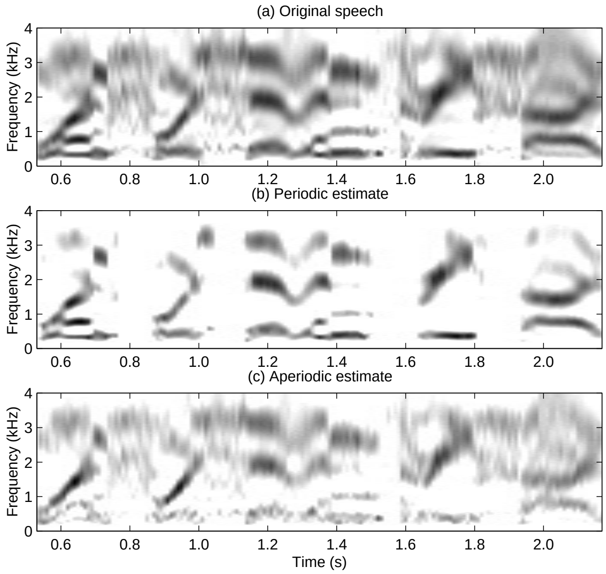

Figure 2: Spectrograms derived from MFCCs of "one-four-seven-three-five" spoken by a female: (a) s, (b) $\hat{v}$, and (c) $\hat{u}$.

Parm.	Front-end processing
base:	→ MFCC → +Δ, +ΔΔ →
split:	→ PSHF ⇉ MFCC ⇉ +Δ, +ΔΔ ⇉ cat →
split1:	→ PSHF ⇉ MFCC ⇉ +Δ, +ΔΔ ⇉ cat →
pca26:	→ PSHF ⇉ MFCC ⇉ cat → PCA → +Δ, +ΔΔ →
pca78:	→ PSHF ⇉ MFCC ⇉ +Δ, +ΔΔ ⇉ cat → PCA →
pca13:	→ PSHF ⇉ MFCC ⇉ PCA ⇉ +Δ, +ΔΔ ⇉ cat →
pca39:	→ PSHF ⇉ MFCC ⇉ +Δ, +ΔΔ ⇉ PCA ⇉ cat →

Table 1: Front-end parameterisations, where "$+\Delta, +\Delta\Delta$" denotes calculation of 1st- and 2nd-order differences, and "cat" implies concatenation of the periodic and aperiodic feature streams. PCA-based parameterisations are distinguished by the size of their analysis matrix, depending on operation order. Thus all parameterisations yield feature vectors with 78 coefficients, except BASE with 39. Note that SPLIT and SPLIT1 differ only in the stream weights used during training.

2.2. 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.,

[1] Adding dither avoided any numerical instabilities that can be induced by training probability distributions with total silence.

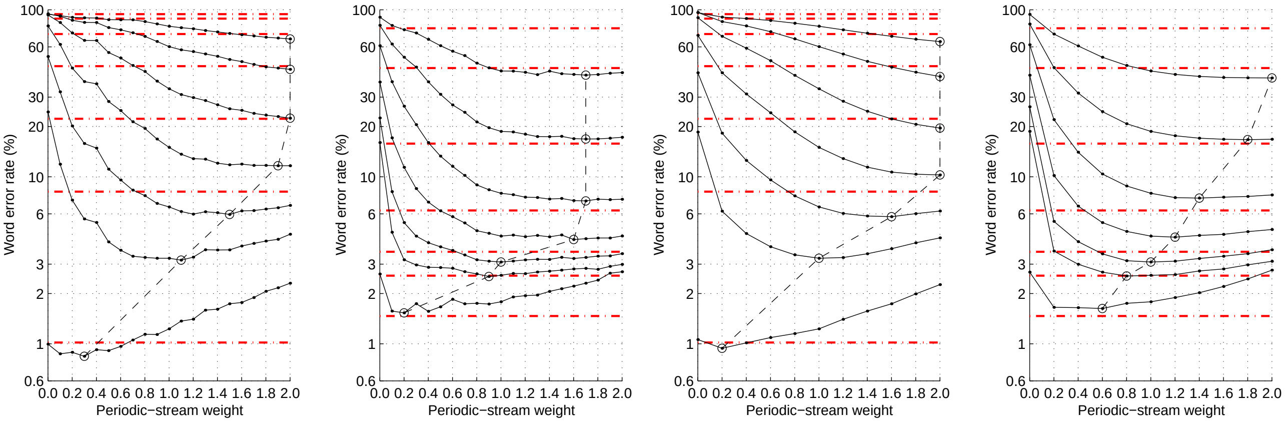

Figure 3: WER (%) for SPLIT (left pair) and SPLIT1 (right pair) versus periodic-stream weight γ_P, with (left) clean and (right) multi-condition training, averaged across each SNR: (from top): -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and ∞ test conditions. Thick dot-dashed lines indicate baseline average scores, and the thinner dashed line (with $\odot$) marks the best at each noise level.

channel distortion). Hence, a recognizer may be trained using only clean data or multiple SNR conditions, and the word error rate (WER) viewed according to test SNR.

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 and SPLIT1 tests, likelihoods of the two streams were weighted independently.

3. Results

3.1. Effects of stream weights

Points for equally-weighted streams, $\gamma_P = \gamma_A = 1.0$ (at the centre of each graph in figure 3), correspond to simple concatenation of the periodic and aperiodic features. The improvement in recognition word accuracy is considerable, especially under noisy test conditions, suggesting that useful information had been masked in the original speech features.

In the SPLIT experiment, the periodic-stream weight was increased from zero to two, in steps of one tenth, and the sum of the weights was held constant. The recognition scores with this changing balance of the streams weights are shown in figure 3 (left pair), and define 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 occurred with all the weight given to the periodic stream; (iii) at intermediate noise levels, a combination of both streams gave the best results. This behaviour was caused by the fact that the PSHF ascribes corrupting noise mainly to the aperiodic component.

Results of the experiments summarised above have been reported elsewhere [10, 11], for which all details of the front-end processing and Viterbi alignment were identical in testing as in training. Now, we consider the case where the weights for the periodic and aperiodic streams, γ_P and γ_A respectively, were held constant and set equal to one for training, but varied as before during the testing stage. The outcome of these recognition tests is shown in figure 3 (right pair). The results indicate a slight degradation in performance at high SNRs, and a modest improvement otherwise. The absence of a significant decrease in performance means that only one stream weight value need be used in training, though it could potentially be varied dynamically when the recognition system was operation.

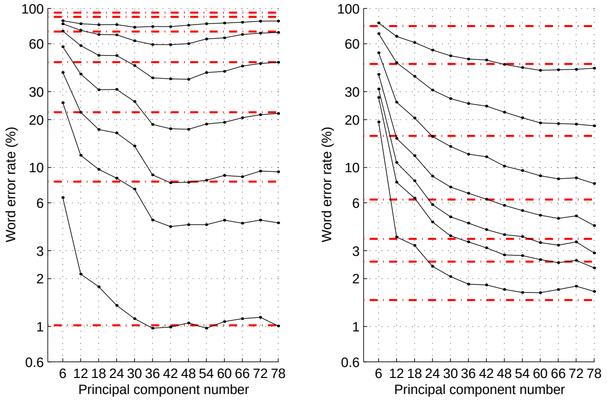

Figure 4: WER (%) for PCA26 versus the total number of PCs, with (left) clean and (right) multi-condition training, as fig. 3.

3.2. Principal component analysis

PCA was used to decorrelate the dimensions of feature data, and sort them by the proportion of variance each dimension explained, from which it could be seen what proportion of the variation in the data was useful to the recognizer. How complementary or redundant the periodic and aperiodic streams were could hence be estimated from the number of principal components (PCs) that were beneficial to the recognition task.

There were generally thirteen dominant dimensions in the data (including the deltas and delta-deltas), but the detection of voiced segments introduced one extra to the periodic component. 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 respectively, and 15 for the recombined streams after concatenation. If the periodic and aperiodic streams were completely redundant, the number of PCs after recombination would be equal to those for the original stream (i.e., 13); if totally independent, the number should be their sum (i.e., 23). As the number of 15 recombined PCs lies between

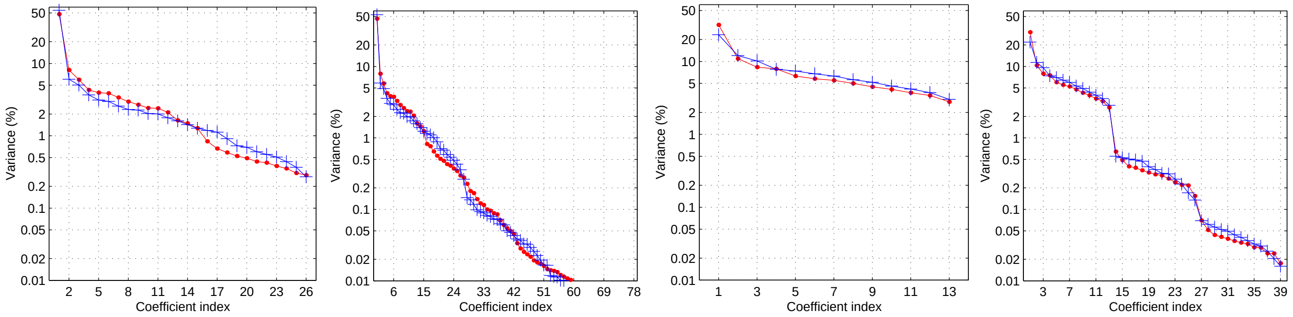

Figure 5: Proportion of variance (%) versus principal component: (· clean, + multi, from left) PCA26, PCA78, PCA13 and PCA39.

13 and 23, it implies that complementary information was contained in the streams separated through the decomposition. Figure 4 shows performance results for the recombined case, as the number of PCs was reduced: first it improved (in the clean training and test cases), then it deteriorated. Figure 5 shows the analysis of variance for the four different PCA parameterisations, and table 2 summarises the best average WERs (minima taking clean and multi averages separately.

	Clean		Multi		Ov-
Parm.	set (a), (b), (c)	Ave.	set (a), (b), (c)	Ave.	erall
base	47.6, 50.4, 41.1	47.4	21.3, 21.1, 23.8	21.7	34.6
split	23.7, 20.3, 25.2	22.6	11.4, 10.1, 12.2	11.0	16.8
split1	22.0, 19.2, 22.8	21.0	11.2, 10.2, 11.9	10.9	16.0
pca26	30.8, 26.1, 31.2	29.0	11.9, 10.4, 12.3	11.4	20.2
pca78	39.6, 35.6, 40.9	38.3	12.4, 11.1, 13.5	12.1	25.2
pca13	28.8, 26.4, 27.9	27.6	13.3, 11.6, 13.5	12.6	20.1
pca39	30.3, 27.1, 31.7	29.3	12.9, 11.4, 14.0	12.5	20.9

Table 2: Best averaged WERs (%) achieved by each front end in table 1, with clean and multi-condition training, and Aurora sets: (a) matched noise and channel, (b) matched channel and unmatched noise, (c) matched/unmatched noise and unmatched channel. The SPLIT and SPLIT1, and PCA results depend respectively on the stream weights and number of selected PCs.

4. Conclusion

The PSHF was used to split each speech waveform in the Aurora 2.0 database into two synchronous streams, periodic and aperiodic, that act 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 connected-digit recognition accuracy scores for various parameterised combinations of the streams, against a conventional one (39 MFCCs, $+\Delta, +\Delta\Delta$). Comparison of the results from using matched weights during training showed a small improvement of approximately 10 % relative to unmatched ones, under clean test conditions. PCA demonstrated augmentation from their combination, but also redundancy between streams. Analysis of the covariation amongst periodic and aperiodic features showed that only 45 (51) of the 78 coefficients accounted for 99 % of the variance, for clean (multi-condition) training, which yielded an 18.4 % (10.3 %) absolute increase in accuracy with respect to the baseline. 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. Further details of this research can be found in Moreno's thesis [10]. 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 aperiodic information. These promising results indicate that it may be worthwhile to investigate applying different forms of front-end processing to each stream, and to consider other forms of model combination, such as in [6] and [12].

5. References

[1] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," *Proc. IEEE-ICASSP*, Minneapolis, MN, vol. 93, no. 2, pp. 550–553, 1993.

[2] C. D. Yoo and J. S. Lim, "Speech enhancement based on the generalised dual excitation model with adaptive analysis window," in *Proc. IEEE-ICASSP*, Detroit, MI, 1995, pp. 832–835.

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

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

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

[6] 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.

[7] M. J. F. Gales, "Maximum likelihood multiple subspace projections for hidden Markov models," *IEEE Trans. on Spch. & Aud. Proc.*, vol. 10, no. 2, pp. 37–47, 2002.

[8] 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.

[9] 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.

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

[11] D. M. Moreno, P. J. B. Jackson, J. Hernando, and M. J. Russell, "Improved ASR in noise using harmonic decomposition," in *Proc. ICPhS*, Barcelona, 2003.

[12] H. J. Nock and S. J. Young, "A comparison of exact and approximate algorithms for decoding and training loosely-coupled HMMs," *Proc. Inst. of Acoust.*, Stratford-upon-Avon, UK, vol. 23, no. 3, pp. 47–60, 2001.