# Biologically inspired features used for robust phoneme recognition

Mitar Milacic*

Griffith School of Engineering,
Griffith University,
Nathan, Qld. 4111, Australia
E-mail: mitar.milacic@griffithuni.edu.au
*Corresponding author

A.P. James

School of CS and IT,
IIITM-Kerala,
IIITM-K Building, Technopark Campus,
Trivandrum, Kerala 695581, India
E-mail: apj@ieee.org

Sima Dimitrijev

Griffith School of Engineering,
Griffith University,
Nathan, Qld. 4111, Australia
E-mail: s.dimitrijev@griffith.edu.au

**Abstract:** Formants are regarded as the basic building blocks of vowels; however, they are very rarely used as features for difficult automatic speech recognition tasks. Formant-based research is generally focused on formant extraction, because of the assumption that a better formant extraction method is the only manner to increase the effectiveness of formants. We challenge this assumption by investigating a different use of formants following their extraction. By using the same principles of combining formants as observed in speech perception studies, we create features that show good recognition performance under noisy testing conditions. Improved recognition performance with the proposed formant features is demonstrated by comparing to Mel-frequency cepstrum coefficients and perceptual linear predictive coding features on a hidden Markov model-based automatic speech recognition system.

**Keywords:** robust speech recognition; formants; signal processing.

**Reference** to this paper should be made as follows: Milacic, M., James, A.P. and Dimitrijev, S. (2013) 'Biologically inspired features used for robust phoneme recognition', *Int. J. Machine Intelligence and Sensory Signal Processing*, Vol. 1, No. 1, pp.46–54.

**Biographical notes:** Mitar Milacic graduated with Bachelor of Engineering in Microelectronic Engineering from Griffith University (hons.) in 2008. He is currently a PhD candidate, with research interests in robust speech recognition.

A.P. James received his PhD from Griffith University, Australia working on a threshold-based face/image recognition method. He is currently the Group Head of the Intelligent Machines and Neuromorphic Engineering Laboratory, Indian Institute of Information Technology and Management, Kerala. His main research interests are in human face recognition, biometrics, image processing, neural circuits, automatic learning and vision, and pattern recognition.

Sima Dimitrijev received his BEng, MSci, and PhD in Electronic Engineering from the University of Nis, Nis, Yugoslavia, in 1982, 1985, and 1989, respectively. He is currently Professor at the Griffith School of Engineering and Deputy Director of Queensland Micro- and Nanotechnology Centre. He is the author of *Principles of Semiconductor Devices*, 2nd ed. (2011, Oxford University Press, New York) and a member of the Editorial Board of *Microelectronics Reliability*.

## 1 Introduction

The most commonly used features in automatic speech recognition (ASR) systems are very sensitive to noise. Even small levels of noise have a large impact on the accuracy of speech recognition experiments. Several studies published in the past decade have focused on addressing this very issue. The method most commonly used to overcome the sensitivity to noise is to model and predict the noise in the noisy speech signal so that it can be removed. However, the model-and-predict approaches often are not able to detect the noise accurately resulting in the loss of useful information (Hermus et al., 2007). This approach is not in line with what we know about how humans hear and perceive speech in noisy environments.

Humans are able to tolerate a lot more noise before recognition accuracy deteriorates and do not use such forms of speech enhancement at the sensory processing level. The ability of humans to concentrate on the speech signal while ignoring the noise content questions the very logic of including noise removal in ASR systems. We recognise this important and practical aspect in our study to develop features that are robust to noise that are inspired by human speech perception studies.

Formants are considered important in defining voiced speech signals (Peterson and Barney, 1952). They can be used as the fundamental features in computerised and biological speech production for voiced speech signals. In voiced signals, formants are the basic building blocks of speech and are known to tolerate noise. Several challenges arise when attempting to utilise formants for ASR systems. The greatest challenge is the extraction of the raw formants with a method that is consistently accurate under noisy and clean conditions. Once the formants are extracted, the next challenge is to utilise the raw formants in the most advantageous way. We focus on the latter of the two described challenges by combining the formants in a manner that emulates the feature extraction process in humans.

Much of the work on formants has, in recent times, been focused on simpler speech recognition problems – either small speaker dependent databases or on larger databases where the test signal has not been significantly corrupted (Yan et al., 2004; Thomas et al., 2008). Further, this work has been largely limited to the extraction of raw formants with greater accuracy, with little emphasis placed on how to best use the raw formants

following their extraction. This work has resulted in many methods to extract raw formants, which range from very complex methods to very simple methods. The more complex methods were created to overcome the difficulties of raw formant extraction in the presence of noise (Niederjohn and Lahat, 1985; Welling and Ney, 1998). One work that attempts to address this problem is by Holmes et al. (1997). They investigate formant use by combining raw formants with MFCC features and show better recognition accuracies when compared to increasing the number of MFCC features. This is a separate idea from the proposed features in this paper where we combine raw formants with one another to increase recognition accuracies.

In this paper, we examine the effect of combining raw formants to create new and useful features for ASR in a noisy environment on continuous speech signals. This approach differs from previously published works in that it is not a new method for raw formant extraction.

## 2 Proposed features

### *2.1 Background*

The use of formants, for vowel classification can be traced back to the days when the objective was the study of vowels rather than ASR. Peterson and Barney (1952) use formats in the analysis of vowels, where they find that it is possible to relate a plot of $F_1$ versus $F_2$ to create what is referred to as a vowel loop. Ten separate vowels are shown to lie on or in the vowel loop, with minimal overlap between them, thus showing good discrimination for the given dataset.

Ohl and Scheich (1997) show that, although there is no direct mapping in a mammalian brain for $F_1$ versus $F_2$, it was possible to directly map $F_1$ versus $F_2 - F_1$ in the auditory cortex of a bat. The auditory cortex of most mammals is similar, which makes this study very relevant for understanding speech recognition in humans. Their comparison of $F_1$ and $F_2 - F_1$ also showed that such features have a good discrimination ability for the given vowels.

Tanji et al. (2003) performed a human listening experiment using Japanese vowels. The experiment was performed on a brain-damaged patient. The damage was such that it made it difficult for the patient to comprehend spoken language, while preserving his ability to speak, read, and write. They showed, in their study, that $F_1 - (F_2 - F_1)$ correlated much better with the recognition accuracies obtained by the patient than $F_2 - F_1$, suggesting that $2F_1 - F_2$ is a useful feature for recognition of phonemes.

These three papers indicate that raw formants and formant combinations could accurately and effectively discriminate between different vowels. They clearly show that classification of vowels can be achieved using the information present in just the first two formants and their combinations.

### *2.2 Method*

The incoming speech signal is separated into overlapped frames with each frame having a length of 25 ms and an overlap of 10 ms with the neighbouring frames. A short time Fourier transform (STFT) is created by applying a discrete Fourier transform (DFT) on each frame:

$$X_{k,j} = \sum_{n=0}^{N-1} x_j(t) e^{-\frac{2\pi i}{N} kn}, \quad k = 0, 1, 2, \ldots, N-1 \tag{1}$$

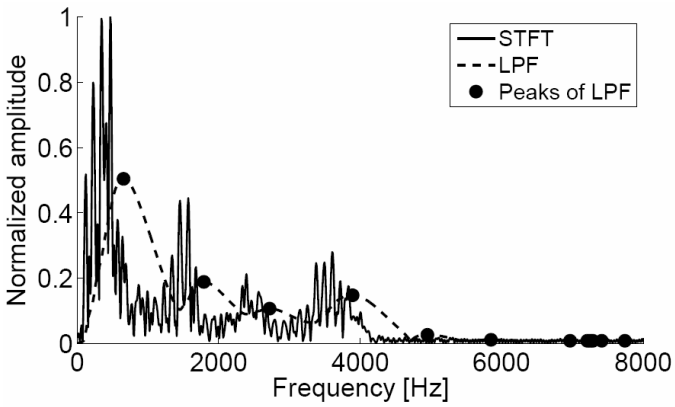
where $N$ is the total number of frequency bins, $j$ is the $j^{\text{th}}$ frame, $t$ is the $t^{\text{th}}$ time sample and $k$ refers to the $k^{\text{th}}$ frequency bin. An example STFT is shown in Figure 1. Each of the STFT frames are passed through a second order low pass Butterworth filter with a normalised cut off frequency of 0.0112, where 1 corresponds to half the sampling frequency. The spectral peaks of the power spectrum generated from the low pass filtered (LPF) signal ($X^{LPF}$) are detected as formants. The peaks of the LPF signal were detected by the following equation,

$$F_h = \begin{cases} M_k, & \left(X_{k,j}^{LPF} - X_{k+1,j}^{LPF}\right) + \ldots \\ & \left(X_{k,j}^{LPF} - X_{k-1,j}^{LPF}\right) > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\theta$ is a threshold used to define the peak, $k$ is the $k^{\text{th}}$ frequency bin, $M_k$ is the frequency that corresponds to bin $k$, $F$ is an array of peaks, and $h$ is an index that is increased by 1 whenever $F_h \neq 0$ and indicates the formant number.

Figure 1 shows the location of formants as obtained by applying the process described using Fourier transform, low pass filtering and peak detection. A shift is present in the LPF signal, due to the length of the filter, however, as it is a constant shift its impact is negligible. Only the first two extracted formants are used as it has been shown that the first and second formants are sufficient for the recognition of vowels.

**Figure 1** A graphical illustration of formant detection as the peaks of an output obtained from a low pass filter (LPF)




Peterson and Barney (1952) suggest that the first two formants, $F_1$ and $F_2$, can be used as features for vowel classification. Ohl and Scheich (1997) give evidence that the features of $F_1$ and $F_2 - F_1$ are represented in the mammalian auditory cortex, suggesting that $F_2 - F_1$ gives a better interpretation of what is occurring inside humans than $F_2$ on its own. Tanji et al. (2003) showed in a human listening experiment on Japanese vowels that $F_1 - (F_2 - F_1)$ correlates much better with recognition accuracies obtained by the patient than $F_2 - F_1$, suggesting that $2F_1 - F_2$ is a useful feature for recognition of phonemes. From a mathematical perspective each feature combination can be looked upon as the

weighted addition of formants $F_1$ and $F_2$. Therefore, in addition to these features, we also propose $F_1 + F_2$ as the simplest formant combination feature with weights equal to one. It should be noted that although these combinations of features have been suggested as cues to speech perception, they were not tested or attempted in ASR environments prior to their use in this paper.

## 3 Experimental setup

### *3.1 Database*

The phoneme recognition task was performed using the TIMIT database (Fisher et al., 1986). It is a standard practice to remove the *sa1* and *sa2* sentences because their inclusion can bias the performance of context dependent systems. The total number of uniquely labelled phoneme classes in the TIMIT database is 61, however, they can be combined together by folding (Lee and Hon, 1989), giving a total of 39 classes for recognition. The database consists of 630 speakers, speaking American English in eight dialects, with each speaker reading ten phonetically rich sentences. The recordings are sampled at 16 kHz.

In order to test the recognition results under noisy conditions, artificially created white Gaussian noise was added in the time domain. To remove the possibility of bias due to changing noise conditions, the corrupted signals were saved to files so that the same noise conditions could be tested on different methods.

### *3.2 Setup*

The recognition experiments were performed using the HTK framework (Young et al., 2006). The MFCC and PLP results reported were obtained by using the HTK HCopy software to create the features. The parameters of both MFCC and PLP features were set so that a Hamming window length of 25 ms was applied, with a window shift of 10 ms. Pre-emphasis was applied to each frame, using a coefficient of 0.97.

For MFCC and PLP features, the results with five and 12 features were obtained by varying the number of output coefficients. The MFCC and PLP results reported using 39 features were obtained by using 12 MFCC or PLP features plus the energy of each frame, the first and second derivatives of these 13 features were calculated, giving a total of 39 features. The performance of the proposed features is compared to the standard mentioned above. We propose the use of a total of five features for recognition:

- $F_1$
- $F_2$
- $F_2 - F_1$
- $2F_1 - F_2$
- $F_1 + F_2$.

## 4 Experimental results

Figure 2 shows that the proposed formant features outperform both PLP and MFCC features when only five features are considered under noisy speech recognition conditions. The speech recognition problem shown is difficult as there is a wide variety of white Gaussian noise being added to the test speech signals. The average recognition accuracies for the three tested SNRs are given in Table 1. The proposed biologically inspired features correspond to higher recognition accuracies than MFCCs or PLPs. This result demonstrates that the formant features represent important information that is useful for ASR and that is robust to varying noise levels in the speech.

**Figure 2** Comparison of recognition accuracies of the five proposed formant-based features with five MFCC features and five PLP features

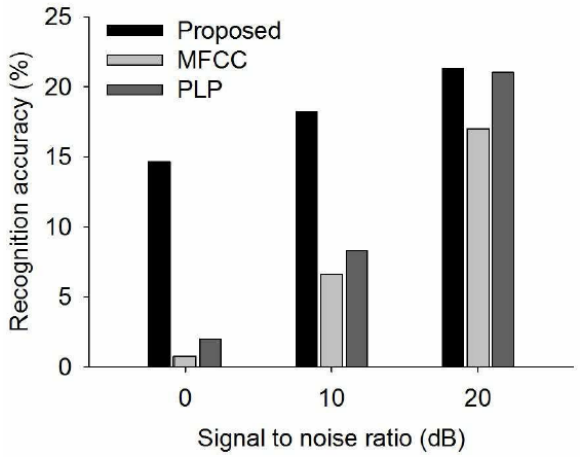



**Table 1** Average recognition performance with five features in noisy conditions

| *Features* | *Recognition accuracy (%)* |
|---|---|
| Proposed | 18.09 |
| PLP | 10.79 |
| MFCC | 8.20 |

Table 2 shows that in very noisy conditions (0 dB) the five formant features outperform the 39 MFCC and PLP feature sets. The performance of the ASR system implemented using the proposed features is less sensitive to the presence of noise than the MFCC and PLP features. When the 0 dB and the clean recognition accuracies are compared, the proposed features have a drop of 13.34%, while a drop in accuracy of 34.53% and 34.50% is seen for five MFCC and five PLP features, respectively. As the number of MFCC and PLP features increases the drop in recognition accuracy increases.

**Table 2** Phoneme recognition results on the entire TIMIT speech database

| *SNR (dB)* | *Recognition accuracy(%)* | | | | | | |
|---|---|---|---|---|---|---|---|
| | *MFCC* | | | *PLP* | | | *Formants* |
| | *5 features* | *12 features* | *39 features* | *5 features* | *12 features* | *39 features* | *5 features* |
| 0 | 0.78 | 1.84 | 11.03 | 2.01 | 3.06 | 6.78 | 14.68 |
| 10 | 6.65 | 9.34 | 29.30 | 8.32 | 11.49 | 23.95 | 18.24 |
| 20 | 17.04 | 26.18 | 51.23 | 21.04 | 28.27 | 46.41 | 21.34 |
| Clean | 35.31 | 48.72 | 71.26 | 36.51 | 49.46 | 70.95 | 28.02 |

As was shown by Holmes et al. (1997), it is possible to improve the performance of PLP and MFCC-based speech recognition systems by combining them with raw formants. In Table 3, we show that this is also the case for our proposed features. Table 3 shows the performance improvement when the proposed formant features are added to the first seven MFCC and seven PLP features. It can be seen, that by combining the proposed features with MFCC or PLP features, there is an improvement in recognition performance under noisy conditions. The recognition performance under clean conditions does not change with the addition of formant features, demonstrating that the addition of formant combinations is useful mostly in noisy test conditions. Further, it should be noted that clean conditions rarely occur in realistic speech environments and the speech recognition results in clean conditions are generally not a good measure of system performance.

**Table 3** Phoneme recognition results for combined features

| *SNR (dB)* | *Recognition accuracy(%)* | | | |
|---|---|---|---|---|
| | *7 MFCC +5 formants* | *12 MFCC* | *7 PLP +5 formants* | *12 PLP* |
| 0 | 5.83 | 1.60 | 2.28 | 3.11 |
| 10 | 15.33 | 7.02 | 15.31 | 11.02 |
| 20 | 27.13 | 20.77 | 27.86 | 27.43 |
| Clean | 46.61 | 46.36 | 47.52 | 48.25 |

Table 2 shows results obtained by the entire TIMIT database, as distinct from the results shown in Table 3 that were obtained on a smaller subset.

## 5 Discussion

ASR under noisy testing conditions remains a challenge even when speech enhancement and noise cancellation methods are utilised. Further, only a noisy speech recognition task reflects a realistic speech recognition problem. We propose and test a premise that biologically inspired features that are based on formants can be more effective and a better match for the human speech recognition processes.

Formants are known as an important cue for speech perception in humans, however, their specific role in ASR is not well understood. The robustness of the formant features

under noisy test conditions confirms that combinations of bio-inspired features are fundamental to speech recognition.

Combining formant features with other features such as MFCC and PLP can improve the robustness of ASR systems. The disadvantage of the formant features is the comparatively lower recognition performance in clean conditions, which can be overcome by combining features. Further, by combining features, the robustness of the conventional features such as MFCC or PLP under noisy test conditions are shown to improve.

It has to be noted that it is possible that there are many more formant feature combinations that may help to improve the reported results. This topic therefore remains open and requires further studies on speech perceptions in the human brain, and is an area of future research.

## 6 Conclusions

In conclusion, it is demonstrated that the use of formant-based features, which are inspired from studies in speech perception, shows robust ASR performance under noisy speech conditions. The ASR system implemented with as few as five formant-based features outperforms the conventional features such as MFCC and PLP when noise levels are high. Using the proposed features along with MFCC or PLP features improves the overall performance of both MFCC- and PLP-based ASR systems.

## References

Fisher, W.M., Doddington, G.R. and Goudie-Marshall, K.M. (1986) 'The DARPA speech recognition research database: specifications and status', *DARPA Workshop on Speech Recognition*, pp.93–99.

Hermus, K., Wanbacq, P. and Hamme, H.V. (2007) 'A review of signal subspace speech enhancement and its application to noise robust speech recognition', *EURASIP Journal on Applied Signal Processing*, No. 1, p.195.

Holmes, J.N., Holmes, W.J. and Garner, P.N. (1997) 'Using formant frequencies in speech recognition', *Proc. Eur. Conf. Speech Communications and Technology (EUROSPEECH)*, pp.2083–2086.

Lee, K-F. and Hon, H-W. (1989) 'Speaker-independent phone recognition using hidden Markov models', *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 11, pp.1641–1648.

Niederjohn, R. and Lahat, M. (1985) 'A zero-crossing consistency method for formant tracking of voiced speech in high noise levels', *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 33, No. 2, pp.349–355.

Ohl, F.W. and Scheich, H. (1997) 'Orderly cortical representation of vowels based on formant interaction', *Proc. Natl. Acad. Sci.*, USA, Vol. 94, pp.9440–9444.

Peterson, G.E. and Barney, H.L. (1952) 'Control methods used in a study of the vowels', *J. Acoust. Soc. Am.*, Vol. 24, No. 2, pp.175–184.

Tanji, K., Suzuki, K., Okuda, J., Shimizu, H., Seki, H., Kimura, I., Endo, K., Hirayama, K., Fujii, T. and Yamadori, A. (2003) 'Formant interaction as a cue to vowel perception: a case report', *Neurocase: The Neural Basis of Cognition*, Vol. 9, No. 4, pp.350–355.

Thomas, S., Ganapathy, S. and Hermansky, H. (2008) 'Hilbert envelope based spectro-temporal features for phoneme recognition in telephone speech', *Interspeech 2008: 9th Annual Conference of the International Speech Communication Association 2008*, Vols. 1–5, pp.1521–1524.

Welling, L. and Ney, H. (1998) 'Formant estimation for speech recognition', *IEEE Transactions on Speech and Audio Processing*, Vol. 6, No. 1, pp.36–48.

Yan, Q., Zavarehei, E., Vaseghi, S. and Rentzos, D. (2004) 'A formant tracking LP model for speech processing in car/train noise', *Proc. ICSLP*.

Young, S.J., Kershaw, D., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (2006) *The HTK Book Version 3.4*, Cambridge University Press, Cambridge.