PHASE BASED FEATURES FOR COGNITIVE LOAD MEASUREMENT SYSTEM

Tet Fei Yap[1,2], Eliathamby Ambikairajah[1,2], Eric Choi[2], Fang Chen[2]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia
tetfei.yap@nicta.com.au, ambi@ee.unsw.edu.au, eric.choi@nicta.com.au, fang.chen@nicta.com.au

ABSTRACT

The current automatic cognitive load measurement system based on MFCC and prosodic features does not take into account phase based speech information. This paper aims to improve the performance of the baseline system by introducing phase based features into the system. The additional features proposed are group delay features, all-pole model based FM features and zero crossing count based FM features. Decrease in performance is observed when phase based features are considered individually or when concatenated with baseline features. However, significant performance improvement is observed when group delay features are fused with baseline features using linear combination score level fusion.

Index Terms— feature extraction, speech classification, cognitive load, group delay, frequency modulation

1. INTRODUCTION

Cognitive load refers to the amount of mental demand imposed on a learner's cognitive system when performing a particular task [1]. Central to the cognitive load theory is the claim that human working memory is limited. Thus, tasks need to be structured in such a way that the load on the human working memory is kept to a minimum. This has important consequences in the field of educational psychology where cognitive load theory is used to design and structure learning tasks to allow for more effective learning [1]. Similarly, cognitive load theory plays a significant role in the design of human-computer interface systems, especially systems which involve large amount of information (for example, a train traffic control system [2]).

When designing a system with the aim of reducing cognitive load requirement, the ability to measure the cognitive load of individuals is crucial. Numerous methods have been proposed in order to measure cognitive load. These methods can be divided into subjective rating techniques, physiological methods, and performance-based measures [1].

Recently, speech features have been identified as a potentially non-intrusive and inexpensive method of measuring cognitive load. Sentence fragments and articulation rate were proposed by Berthold et al. as a method of assessing a user's cognitive load level [3]. A semi-automatic system was designed by Muller et al. which utilizes a wide range of features such as silent pauses and disfluencies [4]. In [5], Yin et al. proposed the use of rate of pauses and rate of pitch peaks as features for an automatic cognitive load measurement system. In [6] and [7], Yin et al. implemented the automatic measurement system using Mel frequency cepstral coefficients (MFCC) and prosodic features together with a Gaussian mixture model (GMM) classifier.

However, the speech features used in [6] and [7] do not take into account phase based information of the speech spectrum. Alsteris and Paliwal have suggested that phase information contains important information that can be utilized in different areas of speech processing [8]. This has already been proven in areas such as emotion detection [9] and speaker recognition [10] where the inclusion of phase related information improves the accuracy of the recognition system.

In this paper, three different phase based features will be introduced as potential features for cognitive load level classification. The performance of these features will then be compared and fused with the baseline system proposed in [7] with the aim of improving the overall performance of the cognitive load measurement system.

2. COGNITIVE LOAD MEASUREMENT SYSTEM

2.1. Baseline System

The baseline system used for comparison of performance is based on the system proposed by Yin et al. [7]. This system utilizes MFCC and prosodic features (pitch and intensity) as the main feature set. These features are later passed through a voice activity detector so that only the voiced region of speech is considered.

Shifted delta coefficients (SDC) are then calculated based on the features to model the long term temporal variation of speech. To account for channel and speaker variability, cepstral mean subtraction (CMS) is applied on the cepstral coefficients and feature warping is performed on each of the feature coefficients [7].

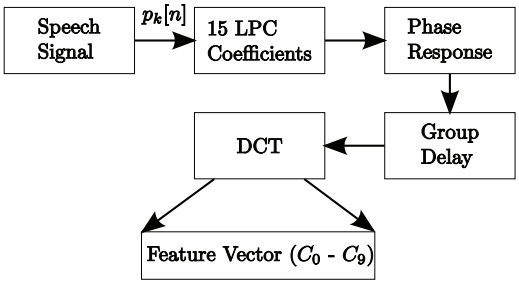

Fig. 1: Extraction of group delay features [9]

2.2. Group Delay Features

A cognitive load measurement system shares certain characteristics with an emotion detection system. For example, both systems aim to detect load level or emotion by detecting changes in speech patterns. Recently, group delay features have been used successfully for emotion detection [9]. This provides the motivation to apply the same features for cognitive load measurement.

Despite the importance of the phase spectrum, one problem encountered by phase features is that phase needs to be unwrapped before it can be used [8].

Group delay, on the other hand, is closely related to phase but does not suffer from the unwrapping problem. Thus, group delay can potentially be used to represent the phase of a speech frame. Formally, group delay is defined as

$$G(f) = -\frac{d\phi(f)}{df} \tag{1}$$

where $\phi(f)$ is the phase of the Fourier transform of the signal $x(t)$. Figure 1 shows the process of extracting the group delay feature vector from a speech signal as proposed by Sethu et al. [9]. In order to estimate the group delay, the all-pole filter coefficients are obtained from the LPC performed on a 10ms speech. The phase response of the filter is then passed through a digital differentiator to obtain the group delay. DCT is applied on the group delay vector to decorrelate and reduce the dimensionality of the feature vector. The first 10 DCT coefficients are used as the group delay feature vector. SDC and feature warping can then be applied on the feature vector.

2.3. FM Features Based on the All-Pole Model

Features based on the AM-FM speech model have been used successfully in the area of speech recognition [11] and speaker recognition [10]. However, such features have yet to be applied to cognitive load measurement.

In the AM-FM model, speech signal $s[n]$ is modeled as the sum of AM-FM signals which corresponds to the vocal tract resonances

$$s[n] = \sum_{k=1}^{K} a_k[n] cos(\phi_k[n]) \tag{2}$$

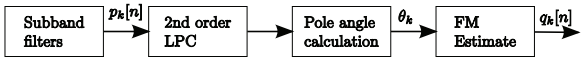

Fig. 2: Estimation of FM components based on the all-pole model

where K is the total number of resonances, $a_k[n]$ is the AM component and $\phi_k[n]$ is the phase of the kth resonance at time index n [10].

Each of the resonances can be isolated using a bandpass filter. The kth bandpass filter output $p_k[n]$ can be represented as

$$p_k[n] = a_k[n] cos\left[\frac{2\pi n f_{ck}}{f_s} + \frac{2\pi}{f_s}\sum_{r=1}^{n} q_k[r]\right] \tag{3}$$

where $q_k[n]$ is the FM component, f_s is the sampling frequency and f_{ck} is the center frequency of the kth band pass filter [10].

Figure 2 describes the method of estimating the FM component $q_k[n]$ by modeling the component using a second order all-pole resonator as proposed by Thiruvaran et al. [10]. The pole angle of the resonator θ_k from the origin is calculated from the second order linear prediction coefficients (LPC). The pole angle is then used to estimate the FM component

$$q_k[n] = \theta_k \frac{f_s}{2\pi} - f_{ck} \tag{4}$$

The FM estimate from each subband is then concatenated to form the FM feature vector. SDC and feature warping can then be applied on the feature vector. This method of FM extraction is reported to produce more reliable FM estimates compared to other techniques such as DESA or Hilbert-transform based methods [10].

2.4. FM Features Based on Zero Crossing Count (ZCC)

One disadvantage of the FM feature extraction technique described in Section 2.3 is the high computational complexity in the overall implementation. In this section, a novel FM feature extraction method with low computational complexity is proposed.

This FM extraction technique is based on the zero crossing count of speech. Consider a continuous speech signal $s(t)$ of frame length l. The signal is first decomposed into smaller bands using a subband filter bank. This is necessary to limit the instantaneous frequency into a small range. For each subband signal $s_k(t)$, the zero crossing count z_k is calculated. The number of full oscillations of the signal can then be approximated as

$$x_k = \frac{z_k}{2} \tag{5}$$

If there are $\frac{z_k}{2}$ number of full oscillations in l seconds, then the period of one full oscillation can be approximated as

$$T_k = \frac{2l}{z_k} \tag{6}$$

The instantaneous frequency is then given by

$$f_{ik} = \frac{z_k}{2l} \tag{7}$$

In the case of a discrete signal with a frame length of N samples

$$l = \frac{N-1}{f_s} \tag{8}$$

where f_s is the sampling frequency. Thus, the instantaneous frequency f_i can be rewritten as

$$f_{ik} = \frac{z_k f_s}{2(N-1)} \tag{9}$$

The FM estimate for each subband is then taken to be

$$f_k = f_{ik} - f_{ck} \tag{10}$$

where f_{ck} is the center frequency of the kth subband filter. The FM feature vector is formed by concatenating f_k of each subband for a particular frame of speech. SDC and feature warping can then be applied on the feature vector.

In practice, simply calculating the zero crossing count for a signal might not take into account oscillations which are not centered about zero. Thus, the zero crossing count of the differentiated signal z_{kd} is calculated instead and the average of the two zero crossing count values is taken $\bar{z}_k = \frac{z_k + z_{kd}}{2}$. Equation 9 is then applied to $\bar{z}_k$ instead of z_k.

Since calculation of zero crossing count involves simple subtraction and sign comparison, this feature extraction technique has a much lower computational complexity compared to the all-pole model based method for FM feature extraction.

3. EXPERIMENT

3.1. Task Design

To test the accuracy of the classification system, speech data are collected from a Stroop test whereby different levels of cognitive load are induced. The experimental data and set up are similar to the Stroop test evaluation performed in [7]. In the Stroop test, subjects are given names of colors printed in different colored fonts. If the font color is different from the color name, the word is said to be printed in an incongruent color. Using the Stroop test, three levels of cognitive load are induced. The low cognitive load task requires subjects to read the color names printed in black or congruent colors. The medium cognitive load task requires subjects to name the font color for color words with incongruent colors. Finally, the high cognitive load task is similar to the medium level task except that time constraints are added to it. Each task lasts for about 30 seconds. Apart from the Stroop test, a separate story reading task of duration 90 seconds is recorded by each of the same participants as training data.

The tests were undertaken by 14 random, native English speaking participants. The evaluation performed was a closed-speaker set evaluation, meaning that all 14 speakers that appeared in the evaluation data existed in the training data as well.

Table 1: Percentage accuracy of individual systems based on baseline and proposed features

	Accuracy (%)
Baseline Features	82.9
Group Delay Features	72.6
All-Pole based FM Features	51.2
ZCC based FM Features	45.2

3.2. Performance Evaluation

A Gaussian mixture model (GMM) classifier is used to evaluate the individual and combined cognitive load measurement systems. The optimal number of mixtures used for modeling the GMM is found to be 128. Due to the lack of training data, background model and adaptation are performed to improve the accuracy of the classifier.

To evaluate the performances of individual systems, the GMM classifier is trained separately for each system. The accuracy of the classification is used as a performance measure of the individual systems.

To evaluate the performances of systems with phase based features combined with the baseline system, two different fusion techniques are used: feature level fusion and score level fusion.

Feature level fusion involves combining the features by concatenation. Score level fusion, on the other hand, involves setting up a GMM model for each feature vector and then applying weights to the loglikelihood scores of the individual systems. In this paper, the loglikelihood score is weighted using a simple linear combination of scores:

$$LL_{fused} = \alpha LL_1 + (1-\alpha) LL_2 \tag{11}$$

where α is the weight, LL_i is the loglikelihood score produced by the ith GMM classifier and LL_{fused} represents the fused score. LL_{fused} can then be used to classify the test utterances. This method of fusion allows the amount of contribution from each system to be controlled. Due to the small data set, the search for the optimum weights is performed using a brute force method on the test data set itself.

4. RESULTS

Table 1 shows the performances of the individual systems based on different feature vectors. It can be seen that the baseline performance is the best, with an accuracy rate of 82.9%. The group delay based system comes second with an accuracy rate of 72.6%. Although this performance is worse than the baseline performance, note that the group delay based system

Table 2: Percentage accuracy of baseline system and fused systems

	Feature Level Fusion (%)	Score Level Fusion (%)
Baseline	82.9	82.9
Baseline + Group Delay	75.0	**85.3**
Baseline + All-Pole FM	65.5	82.9
Baseline + ZCC FM	75.4	82.9

has a lower computational complexity since the feature vector consists of only 40 coefficients as compared to the baseline feature vector which consists of 72 coefficients. FM features, on the other hand, performs very poorly when compared to the baseline system.

Table 2 shows the performances of the fused systems compared to the baseline system. In the case of feature level fusion, all three proposed features do not improve the accuracy rate of the baseline system at all. This might be due to the fact that feature level fusion assigns equal weighting to all features in the fused system. This might cause the low accuracy of phase based features to pull down the accuracy of the baseline features when the two feature sets are combined.

When score level fusion is used on the baseline and group delay based system, an improvement of about 2.4% is observed (as shown in Table 2). This improvement can be explained by observing which sound files are classified wrongly by each individual system. The error pattern of the group delay based system is different when compared to the baseline system. When the likelihood scores for the individual systems are weighted during fusion, this translates to an improvement in accuracy. On the other hand, the error patterns of the FM features are a subset of the baseline system. Thus, FM features do not contribute at all in the fused system.

The results show that as an individual system, group delay based system performs worse than the baseline system. However, group delay features carry some phase-related information not contained in the baseline features and this translates to an improvement in accuracy when the features are fused with the baseline system using score level fusion.

5. CONCLUSION

This paper has proposed the use of phase based features to improve the current baseline cognitive load measurement system consisting of MFCC and prosodic features together with shifted delta coefficients. FM based features do not appear to be effective when compared to the baseline system. However, improvement in performance is achieved when linear combination score level fusion is used to combine baseline system with group delay based system. Thus, while better features need to be explored, this result also indicates that score level fusion provides a promising new avenue for future research.

6. REFERENCES

[1] F. Paas, J.E. Tuovinen, H. Tabbers, and P.W.M. Van Gerven, "Cognitive load measurement as a means to advance cognitive load theory," *Educational Psychologist*, vol. 38, no. 1, pp. 63–71, 2003.

[2] B. Sandblad, A.W. Andersson, I. Frej, and A. Gideon, "The role of human-computer interaction in design of new train traffic control systems," in *Proceedings of World Congress on Railway Research*, 1997, pp. 777–783.

[3] André Berthold and Anthony Jameson, "Interpreting symptoms of cognitive load in speech input," in *Proceedings of International Conference on User Modeling*, 1999, pp. 235–244.

[4] C. Muller, B. Grossmann-Hutter, A. Jameson, R. Rummer, and F. Wittig, "Recognizing time pressure and cognitive load on the basis of speech: An experimental study," *Lecture Notes In Computer Science*, pp. 24–33, 2001.

[5] B. Yin and F. Chen, "Towards automatic cognitive load measurement from speech analysis," *Lecture Notes in Computer Science*, vol. 4550, pp. 1011, 2007.

[6] Bo Yin, Natalie Ruiz, Fang Chen, and M. Asif Khawaja, "Automatic cognitive load detection from speech features," in *Proceedings of CHISIG*, 2007, pp. 249–255.

[7] Bo Yin, Fang Chen, Natalie Ruiz, and Eliathamby Ambikairajah, "Speech-based cognitive load monitoring system," in *Proceedings of ICASSP*, 2008, pp. 2041–2044.

[8] Leigh D. Alsteris and Kuldip K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, May 2007.

[9] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps, "Group delay features for emotion detection," in *Proceedings of Interspeech*, 2007, pp. 2273–2276.

[10] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *IEE Electronics Letters*, vol. 44, no. 6, pp. 449–450, Mar. 2008.

[11] D.V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, 2005.