Available online at www.sciencedirect.com

ScienceDirect

Speech Communication xxx (2011) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Investigation of spectral centroid features for cognitive load classification

Phu Ngoc Le [a,b,*], Eliathamby Ambikairajah [a,b], Julien Epps [a,b], Vidhyasaharan Sethu [a], Eric H.C. Choi [b]

[a] School of Electrical Engineering and Telecommunications, The University of New South Wales, UNSW Sydney, NSW 2052, Australia
[b] ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia

Received 6 September 2010; received in revised form 7 January 2011; accepted 9 January 2011

Abstract

Speech is a promising modality for the convenient measurement of cognitive load, and recent years have seen the development of several cognitive load classification systems. Many of these systems have utilised mel frequency cepstral coefficients (MFCC) and prosodic features like pitch and intensity to discriminate between different cognitive load levels. However, the accuracies obtained by these systems are still not high enough to allow for their use outside of laboratory environments. One reason for this might be the imperfect acoustic description of speech provided by MFCCs. Since these features do not characterise the distribution of the spectral energy within subbands, in this paper, we investigate the use of spectral centroid frequency (SCF) and spectral centroid amplitude (SCA) features, applying them to the problem of automatic cognitive load classification. The effect of varying the number of filters and the frequency scale used is also evaluated, in terms of the effectiveness of the resultant spectral centroid features in discriminating between cognitive loads. The results of classification experiments show that the spectral centroid features consistently and significantly outperform a baseline system employing MFCC, pitch, and intensity features. Experimental results reported in this paper indicate that the fusion of an SCF based system with an SCA based system results in a relative reduction in error rate of 39% and 29% for two different cognitive load databases.

Keywords: Cognitive load; Gaussian mixture model; Spectral centroid feature; Frequency scale; Kullback–Leibler distance

1. Introduction

The cognitive load of a person refers to the amount of mental demand imposed on the person when performing a particular task, and has been closely associated with the limitations of human working memory (Shriberg et al., 1992; Paas et al., 2003). Research on cognitive load has shown that when a user is performing a task, performance will degrade if the load level is too low or too high (Paas et al., 2003). In some cases however, it may be possible to adjust the workload of the user and if the cognitive load level of the user can be estimated or classified along an ordinal scale, it may be possible to tailor these adjustments to the workload such that productivity can be improved. In the past 15 years or so, a number of techniques have been proposed to measure the cognitive load level, including techniques based on: physiological measures such as heart rate, brain activity, eye movement (Pass and Merrienboer, 1994; Gerven et al., 2004); behavioural measures such as mouse speed and pressure, linguistic and dialog patterns (Berthold and Jameson, 1999; Muller et al., 2001); performance measures such as reaction time, accuracy, error rate; and self-reported subjective ranking of experienced load

* Corresponding author at: School of Electrical Engineering and Telecommunications, The University of New South Wales, UNSW Sydney, NSW 2052, Australia.

E-mail addresses: phule@unsw.edu.au (P.N. Le), ambi@ee.unsw.edu.au (E. Ambikairajah), j.epps@unsw.edu.au (J. Epps), vidhyasaharan@gmail.com (V. Sethu), Eric.Choi@nicta.com.au (E.H.C. Choi).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.01.005

level on single or multiple rating scales (Paas et al., 2003; Yin et al., 2008).

Among these, behavioural measures based on speech have been recognised as a particularly good choice since speech data exists in many real-life tasks (e.g. telephone conversation, voice control) and can be easily collected in non-intrusive and inexpensive ways. Previous studies have revealed systematic influences of cognitive load on specific aspects of speech such as, disfluencies, articulation rate, content quality, number of syllables, silent pauses, filled pauses (Muller et al., 2001); sentence fragments and articulation rate (Berthold and Jameson, 1999); average pause length, average pause frequency, and average response latency (Khawaja et al., 2007). However, while these high level features can potentially be used for cognitive load level recognition (Berthold and Jameson, 1999; Muller et al., 2001), the extraction of these features relies on either manual labelling of the speech data or automatic speech recognition (ASR), neither of which are very attractive given that manual labelling is too slow and expensive and ASR systems may not yet be robust enough for this application.

These high level features are, however, not the only influences that cognitive load has on speech. It has been shown by Steeneken and Hansen (1999) that among the physiological consequences of the mental workload are respiratory changes, e.g. increased respiration rate, irregular breathing, and increased muscle tension of the vocal cords. The increased muscle contraction of the vocal cord and vocal tract may adversely affect the quality of speech. Consequently, low level speech features that characterise the vocal cords, such as the fundamental frequency (F_0), and those that characterise the shape of the vocal tract, such as formant frequencies, have been found to reflect the cognitive load experienced by the speaker. In particular, increases in cognitive load have been associated with increases in F_0 (Scherer et al., 2002; Mendoza and Carballo, 1998; Griffin and Williams, 1987; Boril et al., 2010), reduction in jitter and shimmer (Mendoza and Carballo, 1998), increases in the first and fourth formants (Boril et al., 2010) and decreases in the second formant (Yap et al., 2010b). Other vowel-specific variations in formant frequencies with cognitive load have also been reported by Yap et al. (2010b). Apart from F_0 and formant frequencies, low level features characterising the spectral energy distribution have also been found to be indicative of cognitive load. An increase in cognitive load is reflected by an increase in spectral energy spread and spectral centre of gravity (Boril et al., 2010), reduction in the ratio of energy below 500 Hz to energy above it (Scherer et al., 2002) and decrease in the gradient of energy decay (Scherer et al., 2002). It has also been suggested that variability in speech amplitude increases while the speech spectra becomes flatter under high cognitive load conditions (Lively et al., 1993). The existence of these systematic variations of low level speech parameters with cognitive load has formed the basis for a number of automatic cognitive load classification systems.

Particularly, the usefulness of F_0, intensity and MFCCs, has been shown by Yin et al. (2008, 2007) and Boril et al. (2010). It was shown by Yap et al. (2009) that the group delay feature, based on the phase characteristic of the speech spectrum, can be used to provide additional cognitive load information to the MFCCs based system and improve its performance. Furthermore, it was indicated by Yap et al. (2010a) and Le et al. (2010) that the features based on the voice source are useful for cognitive load classification. The usefulness of formant frequencies was also found by Boril et al. (2010) and Yap et al. (2010b, 2010). The non-linear Teager energy operator was found to be effective for classifying cognitive loads by Fernandez and Picard (2003). Other features including perceptual linear prediction coefficients, spectral centre of gravity, spectral energy spread, and the vowel durations were also found to be useful for cognitive load classification systems as outlined in (Boril et al., 2010). Further analyses of speech-based features can be found in (Le et al., 2009) where the spectral distribution of information pertaining to cognitive load was studied. The results of this study suggest that the amount of cognitive load specific information in the 0–1 kHz frequency band is particularly high.

Although several automatic speech-based cognitive load classification systems have been developed, these systems are still limited in terms of performance. The system reported by Yap et al. (2010a) results in a classification accuracy of 84.4%, which is the highest accuracy among those reported in the literature, involving three levels of cognitive load based on the Stroop test corpus, one of the two corpora used in this study (described in detail in Section 3). Among the speech features that have been used in automatic cognitive load classification systems, MFCCs have been recognised as one of the most effective features when the shift delta coefficients (SDC) of the features are used (Yin et al., 2008, 2007; Yap et al., 2010a; Le et al., 2010). MFCCs are a compact representation of the spectral envelope which reflects vocal source and the vocal tract filter information. The MFCC features capture the shape of the speech spectral envelope based on subband magnitude spectrum estimates obtained using a series of mel scale filters and although effective, they do not completely characterise the spectral envelope. Some details of the speech spectrum, such as the spectral energy distributions within mel filter subbands are not captured by MFCCs due to the limitation that information in each subband is represented by only one value, representing the total spectral energy contained in that subband. Spectral centroid features can be used to capture more information about these subband spectral distributions and have been shown to be effective for speech recognition system (Gajic and Paliwal, 2006; Paliwal, 1998) and speaker recognition system (Hosseinzadeh and Krishnan, 2008). These features have certain similarities to formant frequencies but can be estimated easily and reliably, unlike the formant features (Paliwal, 1998). Also, since features based on formant frequencies have been recognised to be effective for cognitive load

Please cite this article in press as: Le, P.N. et al., Investigation of spectral centroid features for cognitive load classification, Speech Comm. (2011), doi:10.1016/j.specom.2011.01.005

classification (Boril et al., 2010; Yap et al., 2010b, 2010), we may expect that the spectral centroid features will also prove to be useful for cognitive load classification system.

In the study reported in this paper, we investigate the use of spectral centroid frequency, termed spectral subband centroid by Paliwal (1998), and spectral centroid amplitude (Kua et al., 2010), in a cognitive load classification system. The number of filters used during their extraction and their spectral characteristics influence the effectiveness of these two features. In this paper, initially the amount of cognitive load-specific information in each subband is estimated. Following this, the filterbank configuration is empirically investigated based on cognitive load classification experiments in order to determine the best configuration.

2. Subband spectral centroid based features

The spectral centroid frequency (SCF) is an estimate of the ‘centre of gravity’ of the spectrum within each subband. Originally proposed as a feature for speech recognition systems (Paliwal, 1998), it has been reported that SCF is a formant-like feature, as it provides the approximate location of the formant frequencies in the subbands (Paliwal, 1998). In addition to spectral centroid frequency, in this paper, we propose the use of another feature, termed spectral centroid amplitude (SCA), the weighted average magnitude spectrum in the subband.

The proposed spectral features are extracted from framed speech segments as follows. Let $s[n]$, where $n \in [0, N-1]$, represent a speech frame (of length N) in the time domain and let $S[f]$ represent the discrete spectrum of this frame. Then, $S[f]$ can be divided into M subbands using a series of Gabor filters (Kleinschmidt, 2002) whose frequency responses are $W_m[f]$, where $m \in [1, M]$.

Assume that the mth subband has a lowest frequency l_m and highest frequency u_m. Each of the two proposed spectral features can be calculated from $S[f]$ for the mth subband as follows.

The spectral centroid frequency (SCF) is computed as the weighted average frequency for a given subband, where the weights are the normalised energy of each frequency component in that subband, as shown in Eq. (1)

$$\mathrm{SCF}_m = \frac{\sum_{f=l_m}^{u_m} f|W_m[f]S[f]|}{\sum_{f=l_m}^{u_m} |W_m[f]S[f]|}. \tag{1}$$

The final SCF feature vector for each frame is obtained by concatenating all the SCF_m.

The Spectral Centroid Amplitude (SCA) is the weighted average magnitude spectrum in the subband, with the frequency serving as weights, as shown in Eq. (2). Average energy could be computed using Eq. (2) by simply setting those weights to be 1 (Kua et al., 2010)

$$\mathrm{SCA}_m = \frac{\sum_{f=l_m}^{u_m} f|W_m[f]S[f]|}{\sum_{f=l_m}^{u_m} f}. \tag{2}$$

The final SCA feature vector for a speech frame is obtained by taking the Discrete Cosine Transform (DCT) of the log of the vector obtained by concatenating all SCA_m in that frame, in order to reduce the dynamic range and decorrelate the features, and can be expressed as

$$\mathrm{SCA}(k) = \frac{1}{\sqrt{M}}\mu_k \sum_{m=0}^{M-1}[log(\mathrm{SCA}_m)]\left[\cos\left(\frac{\pi}{2M}(2m+1)k\right)\right] \tag{3}$$

where $k = 0, \ldots, M-1$; $\mu_0 = 1$; $\mu_k = \sqrt{2}$ for $1 \leqslant k \leqslant M-1$.

In the case of the SCF, the DCT is not applied, similarly to (Paliwal, 1998) as it is a frequency based feature. Moreover, in (Thiruvaran et al., 2006) the DCT was not applied to the frame-averaged FM feature, which is another frequency based feature similar to the SCF. Henceforth, SCF_m, SCA_m will refer to the feature values in each subband, and SCF, SCA will refer to the final spectral centroid feature vectors.

As previously mentioned, the MFCCs are computed from the total energy in each subband and hence will only reflect variations in the total energy in a subband. However, there are instances where the distribution of energy within each subband varies but the total energy does not, and MFCCs will not reflect this. The use of frequency as weights for computing the SCA_m allows the variation of the spectrum distribution in these instances to be reflected in the SCA_m values, as shown in Fig. 1. It can be observed from this figure that the energies of the two spectra are the same but the SCA_m are different.

As explained, the SCF_m and SCA_m capture different aspects of the spectral distribution in each subband and therefore are expected to complement each other. The complementary nature of these features is illustrated in Fig. 2, which shows the spectral centroid features corresponding to different examples of synthetic spectra comprising of straight lines with varying slopes, in two different subbands. It can be observed that the resultant variations in SCF_m and SCA_m are very different. Moreover, there are regions of the energy-slope plane where one of the two

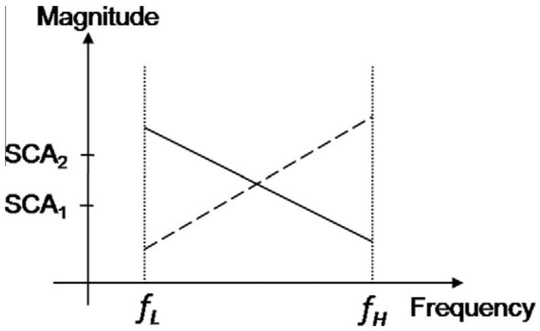

Fig. 1. Example of the spectra in a subband $[f_L, f_H]$, the solid line is the first spectrum and the dashed line is the second spectrum, after (Kua et al., 2010). These spectra have the same energies ($E_1 = E_2$) but different spectral centroid amplitudes ($SCA_1 \neq SCA_2$).

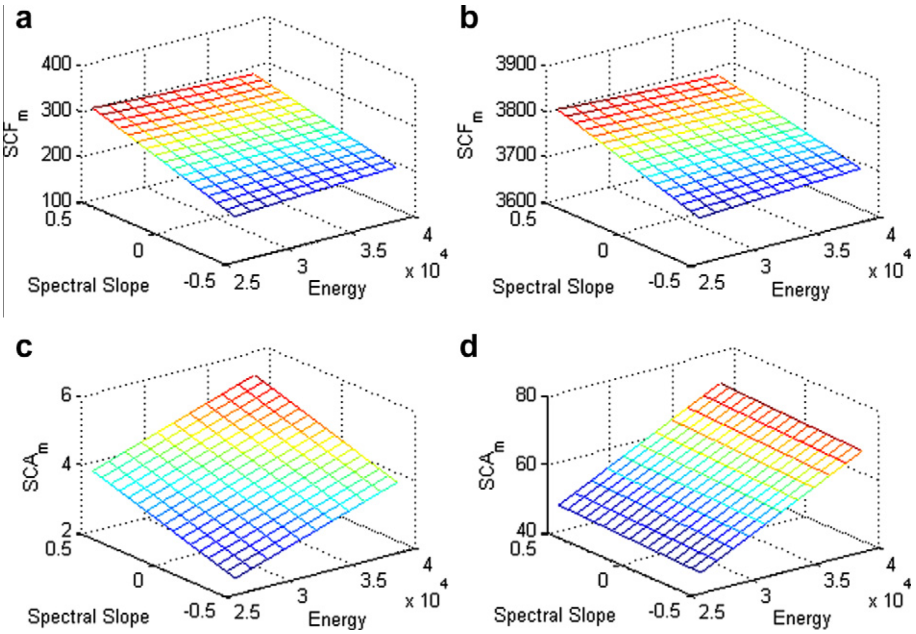

Fig. 2. The variation of the SCF_m and SCA_m in two subbands (a) and (c) for the low frequency subband, and (b) and (d) for the high frequency subband.

features varies more than the other. In Fig. 2, lines of constant energy correspond to constant MFCC feature values (and hence MFCCs cannot distinguish between them), by way of contrast with the SCA_m and SCF_m. The use of the frequency as the weight also results in the SCA_m in different subbands being very different as shown in Fig. 2c and d, although these subbands have the same spectral distributions.

Fig. 3 shows the spectral centroid features as well as the spectral envelopes for three different cognitive load levels (*low, medium,* and *high*) extracted from an utterance of the vowel /*ey*/ by a female speaker in the Stroop test database (Section 3). In this example, the spectral centroid features were extracted by splitting the speech spectrum into six non-overlapping subbands equally spaced in the mel scale (the number of subbands was chosen as six based on preliminary experiments). It can be observed that the roll off in the spectral envelope is steeper for the high cognitive load level. Since the SCA_m are computed as the weighted average amplitude spectrum in each subband, this large negative slope results in the amplitude of the high frequency SCA_m for the high cognitive load being substantially lower than those for the low cognitive load. On the other hand, the low frequency SCA_m for high cognitive load is larger than that for low cognitive load. In addition to the differences in spectral slopes, it can also be observed that the spectral energy distributions in the individual subbands are different, which results in the SCF_m in each subband varying between different cognitive load levels.

Fig. 4a and b shows the statistical spread of the coefficients of the six dimensional SCF and SCA, computed from speech of the vowel /uw/ corresponding to a female speaker in the Stroop test corpus. In these figures, the thick bar extends from the 15th to the 85th percentile, the thin bar extends from the 5th to the 95th percentile, and the middle strip indicates the mean of the distribution. The potential for discrimination between different CL levels can be observed from these figures. Although individual coefficients of these features do not show significant differences between the three CL levels, their combination is more promising, as will be seen.

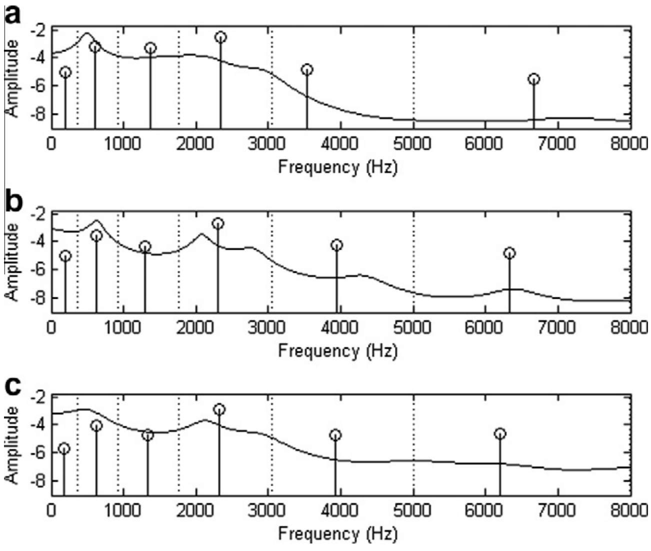

Fig. 3. Subbband spectral centroid frequencies (SCF_m), spectral centroid amplitudes (SCA_m), and linear predictive spectral envelope of the vowel /ey/ of a female speaker for (a) *high cognitive load*, (b) *medium cognitive load*, and (c) *low cognitive load* from the Stroop test database. The SCF_m are shown by the location of the stems, the SCA_m are shown by the amplitude of the stems, the subband boundaries are shown by the dashed vertical lines, and the spectral envelope is shown by the solid continuous curves.

The stages involved in the computation of the spectral centroid features are illustrated in Fig. 5.

Please cite this article in press as: Le, P.N. et al., Investigation of spectral centroid features for cognitive load classification, Speech Comm. (2011), doi:10.1016/j.specom.2011.01.005

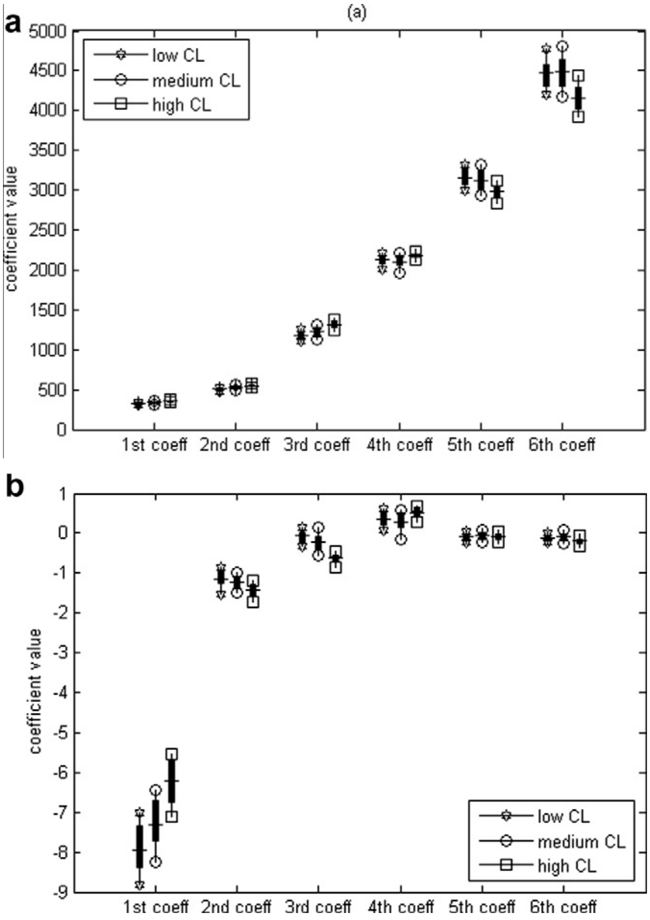

Fig. 4. (a) Statistical variation of the six coefficients of SCF over the three levels of cognitive load speech of the vowel /uw/ of a female speaker in the Stroop test corpus. The thick bar extends from the 15th to the 85th percentile, and the thin bar extends from the 5th to the 95th percentile. The middle strip indicates the mean. (b) Statistical variation of the six coefficients of SCA over the three levels of cognitive load speech of the vowel /uw/ of a female speaker in the Stroop test corpus. The thick bar extends from the 15th to the 85th percentile, and the thin bar extends from the 5th to the 95th percentile. The middle strip indicates the mean.

3. Cognitive load database

All experiments reported in this paper were performed on two databases, referred to herein as the Stroop test and Reading and Comprehension databases (Yin et al., 2008, 2007). Both contain speech corresponding to three cognitive load levels (*low, medium*, and *high*), corresponding to different experimentally induced levels of task difficulty, from 15 native English speakers (eight females and seven males). The speech in both databases is sampled at a rate of 16 kHz.

For the Stroop test corpus, low cognitive load speech was recorded by asking the subjects to read out words (names of colours) written in black or congruent font colour (i.e., colour of the font is the same as the word to be read); speech corresponding to medium cognitive load was recorded by asking the subjects to name the font colour of the words written in incongruent colour (i.e., the font colour of the words is different to the meaning of the colour words); and speech corresponding to high cognitive load was recorded in the same manner as for medium cognitive load speech except that a time constraint was imposed on the subjects when performing the task. Recordings of these tasks contain approximately 60 s of speech (4 utterances, 15 s for each utterance) from each subject, including each of the three cognitive load levels. An additional task was recorded by asking the same participants to read a story with duration approximately 90 s, as neutral reference data.

For the Reading and Comprehension corpus, speech corresponding to each cognitive load level was recorded by asking the subject to read out a story of a corresponding level of difficulty and then answer three open ended questions related to the content of the story. The difficulty levels of the stories were estimated based on the Lexile scale (Metametrics, 2007) – a semantic difficulty and syntactic complexity measure scale ranging from 200 to 1700 Lexiles (L), corresponding to the reading level expected from a first grade student to a graduate student. The Lexile ratings of the stories used were 925 L, 1200 L, and 1350 L, respectively. The speech in this corpus contains four utterances corresponding to each cognitive load level for each subject, one from reading the story and three from answering the questions related to that story. The approximate lengths of the utterances corresponding to reading the story for the low, medium and high cognitive loads are 90 s, 140 s, and 230 s, respectively. The approximate length of each answer to the three questions for all three levels of cognitive load is 30 s.

Comparing the two corpora, the Stroop test corpus contains a limited number of colour words such as ‘red’, ‘blue’, ‘green’, etc. This corpus is akin to an isolated speech corpus as the subjects read the words one by one slowly. In addition, there is a speech rate artifact caused by the time constraint for the high cognitive load speech. On the contrary, the Reading and Comprehension corpus contains a significantly larger vocabulary due to the varied content of the

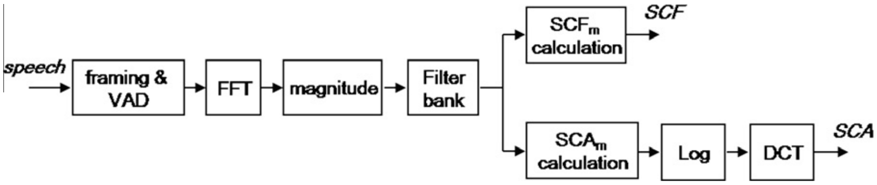

Fig. 5. Block diagram of SCF and SCA feature extraction (Kua et al., 2010).

long stories, and the open ended nature of the questions. The story reading section of this corpus contains continuous speech and the question answering section consists of spontaneous speech. The high level of phonetic variability and short-term cognitive load variability in the Reading and Comprehension corpus results in greater variability in speech features for each load level in this corpus compared to the Stroop test corpus. Consequently, classification of the cognitive load level based on speech from the Reading and Comprehension corpus is more challenging than classification based on speech from the Stroop test corpus.

4. Spectral centroid features – spectral distribution of cognitive load specific information

The spectral centroid based features capture information about the speech spectrum via a bank of filters. In order to efficiently capture information about cognitive load, it is beneficial to determine how this information is distributed across the speech spectrum. The experiments reported in this section aim to determine this distribution of cognitive load specific information contained in SCF_m and SCA_m across different frequency bands empirically by estimating their relative per-band contributions towards cognitive load discrimination. In this experiment, speech was decomposed into 32 subbands using 32 Gabor bandpass filters spaced uniformly across its bandwidth and the SCF_m and SCA_m features were computed in each of these subbands. A uniform filterbank was used in this analysis to avoid the variability caused by the variations in bandwidth that are present in other commonly used filterbanks based on the mel and ERB frequency scales. For each cognitive load level, Gaussian mixture models (GMMs), described in Section 5.2, were then trained using the SCF_m and SCA_m features from each subband individually, thus giving one GMM per cognitive load level per subband. The Kullback–Leibler distance (KL distance), which has been widely used to estimate the separation between two probabilistic models (Le et al., 2009; Goldberger and Aronowitz, 2005), was adopted to estimate the separation between subband specific GMMs corresponding to different cognitive load levels.

Since the aim of the experiment reported in this subsection was to estimate the distribution of cognitive load specific information across the speech spectrum, it is important to use the best possible feature representation prior to modelling using GMMs. Consequently, the original feature vector was concatenated with the shifted delta coefficients (SDCs) of the SCA_m and SCF_m and used to train the GMMs since SDCs have been shown to be effective in capturing temporal variations of spectral features for cognitive load classification (Yin et al., 2008; Le et al., 2009). The pairwise KL distance, was determined in each subband between 256-mixture GMMs corresponding to the different cognitive load levels. An average KL distance in each subband was then obtained by taking the mean of all the pairwise KL distances between GMMs trained from features extracted from that subband. These average KL distances, computed for all 32 subbands for both spectral centroid features (SCA_m and SCF_m) as well as average subband energy (the precursor to the MFCC feature), are shown in Fig. 6. The KL distances reported here were normalised such that the maximum distance was 1. This normalisation was carried out to remove the influence of the magnitudes of the feature values (which are not relevant) on the KL distance.

It is interesting to observe that the distributions of the cognitive load information across different frequency bands are highly consistent between both corpora. For the SCF_m, the frequency band from 250 Hz to 750 Hz is the most significant band for cognitive load classification. Apart from this band, the frequency band from 1250 Hz to 1500Hz is also quite important compared with the other frequency bands. Beyond these two bands, the contribution of SCF decreases steadily until the last frequency band. As SCF_m capture the spectral distribution within subbands, the effectiveness of the SCF_m in the two above-mentioned frequency bands suggests that the variations of spectral distribution in those frequency bands are important to characterise the difference between cognitive loads. These two frequency bands roughly correspond to the first and the second formants (Peterson and Barney, 1952). These results are also consistent with those reported by Yap et al. (2010b) which found that the first two formants could be used to classify cognitive load effectively, and they were more significant than the third formant. It can also be noted that the spectral distributions of cognitive load specific information for both SCA_m and average subband energy are very similar to each other and this is most likely because they capture similar information. The most significant band from the point of view of cognitive load classification, with regards to these two features is around 750–1000 Hz. This is consistent with the results reported by Le et al. (2009) where cepstral coefficients computed in the frequency band from 0 Hz to 1000 Hz were shown to be more important than those computed from other frequency bands.

The large amount of cognitive load specific information distributed in the low frequency bands suggest that it may be advantageous to adopt a frequency scale that provides a higher resolution at lower frequencies for feature extraction.

5. Automatic cognitive load classification system

5.1. Front-end

In an automatic cognitive load classification system based on low level spectral, acoustic and prosodic features, each cognitive load level is typically modelled by a statistical model of these low level speech features, while the cognitive load level is taken as that whose model maximises the

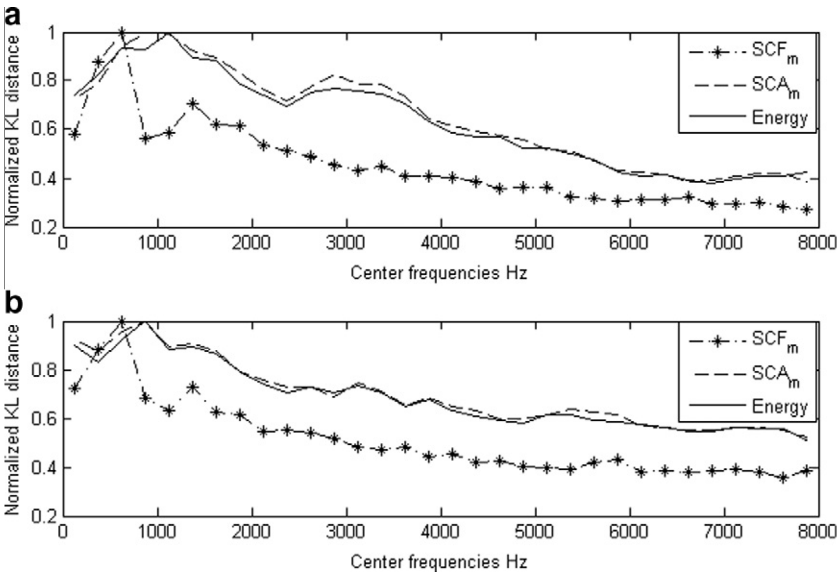

Fig. 6. The normalised Kullback–Leibler distance of SCF_m, SCA_m, and energy in different frequency bands computed across (a) Stroop test corpus and (b) Reading and Comprehension corpus.

likelihood of observing the test speech features (Yin et al., 2008; Le et al., 2010).

All speech features used in this study are extracted from the voiced part of the speech signal using 25 ms frames with a 15 ms overlap between consecutive frames. Voiced frames are determined as those whose pitch can be estimated using the RAPT algorithm (Talkin, 1995). Feature warping (Pelecanos and Sridharan, 2001) is then utilised to map the distribution of the feature vectors extracted from an utterance to a Gaussian distribution of zero mean and unit variance in order to reduce variability caused by differences between different speakers. The shifted delta coefficients (SDC) are extracted to capture information pertaining to the temporal evolution of the static spectral centroid features. It has previously been shown that the concatenation of the static features with dynamic features (SDCs) gives a system that outperforms the one that uses static features only, with regards to systems based on MFCCs and prosodic features (Yin et al., 2008; Le et al., 2010).

5.2. Back-end

A Universal Background Model–Gaussian Mixture Model (UBM–GMM) based classifier (Reynolds et al., 2000) was used as the back-end for the systems reported in this paper. In the UBM–GMM scheme, initially a GMM that models the distribution of the features corresponding to all the data (from all classes) is trained, referred to as the universal background model (UBM). This UBM is then adapted using class-specific data to obtain Gaussian mixture models for each class using the Maximum A Posteriori (MAP) adaptation (Reynolds et al., 2000). Preliminary tests indicated that systems using the UBM–GMM scheme performed better than similar systems (i.e., using the same features) using independently trained GMMs.

All classification experiments reported in this paper were performed in a speaker independent manner, where data from speakers appearing in the test data were not present in the training data. For experiments performed on the Reading and Comprehension corpus, data from a set of five speakers were used as the test dataset and that from two other sets of five speakers each were used as the training set. All experiments were performed three times with a different set of five speakers forming the test set on each occasion and all results reported were obtained by averaging over the three instances.

On the other hand, for experiments performed on the Stroop test corpus, data from one speaker was used in the testing phase and that from the other fourteen speakers were used in the training phase. Each experiment was performed 15 times with a different speaker used as the test speaker each time, and the results were averaged. This mode of experiment was used for the Stroop test corpus previously (Yin et al., 2008; Yap et al., 2010a; Le et al., 2009) and was adopted again for ease of comparison with this work.

For experiments conducted on the Stroop test database, features extracted from the story reading speech, from all subjects allocated to the training set, were used to train the UBM. Features extracted from speech corresponding to each of the three cognitive load levels, from all subjects allocated to the training set, were used to adapt this UBM to obtain the GMMs corresponding to the three cognitive load levels in the training phase. In the testing phase, the features extracted from speech corresponding to each of the three cognitive load levels, from all subjects allocated to the test set, were used to determine likelihood scores from the GMMs. For experiments conducted on the Reading and Comprehension database, features extracted from the story reading speech of all three cognitive load levels, from all subjects allocated to the training set, were used to train the UBM. The features extracted from the question answering speech, from all subjects allocated to the training set, were used to adapt this UBM to obtain the three GMMs in the training phase. In the testing phase, the features extracted from the question answering speech, from all subjects allocated to the test set were used to determine likelihood scores from the GMMs. If only a single feature set was used, the classification result was then obtained as the load level corresponding to the model that best matched the test data (whose likelihood score is maximum). In the case of more than one feature set being used, the log-likelihood scores generated by the classifiers for each of the different feature set were fused using a linear search fusion technique to obtain the final decision.

In particular, the fused log-likelihood score was obtained from the linear combination of log-likelihood scores of individual systems as follows:

$$LL_{fused} = \sum_{i=1}^{N} \alpha_i LL_i, \tag{4}$$

where LL_{fused} is the fused log-likelihood score, LL_i and α_i are the log-likelihood score and the weighting coefficient of the ith system respectively. The weighting coefficient satisfies $0 \leqslant \alpha_i \leqslant 1$ and $\sum_{i=1}^{N}\alpha_i = 1$ and was empirically chosen to optimise the performance of the system. Specifically, in this study the value of α_i was varied from 0 to 1 with the step of 0.01. The chosen value of α_i was the one that produced the highest accuracy for the classification system. The effect of varying these weighting coefficients on the performance of the classification system based on fusing the results of the SCF based system and the SCA based system is further described in Section 6.3.

The number of mixtures used in this study for the UBM and all three GMMs is 256. This number was chosen based on preliminary experiments that suggested this choice offers the highest performance.

5.3. Baseline system

MFCC and prosodic features (pitch and intensity) have been established as some of the most effective features for a speech-based automatic cognitive load classification system. In this paper, a system based on these features is selected as the baseline system and the performance of all other systems are compared to it. Seven MFCCs were extracted from 23 triangular mel scale filters, similar to the feature used by Yin et al. (2008) and Yap et al. (2010a), the pitch contour was extracted using the RAPT algorithm (Talkin, 1995), and the intensity was extracted using Praat software. The input feature of the baseline system was the concatenation of the MFCCs, pitch and intensity.

The classification accuracy of the baseline system performed on the Stroop test corpus reported in Table 1 is comparable with the accuracy of the baseline system reported by Yap et al. (2010a).

Table 1
Performance of system employing baseline features (concatenation of MFCCs, pitch, and intensity).

Classification accuracy (%)	Without dynamic	With dynamic
Database		
Stroop test	58.7	78.9
Reading and Comprehension	42.2	60.7

6. Development of cognitive load classification system using spectral centroid features

6.1. Analysis of number of subbands

In order to investigate the performance of SCF and SCA features extracted with a varying number of subbands, a Gabor filterbank with filters equally spaced on the mel scale was used to extract spectral centroid features, with the number of filters in the array varying from 2 to 22. In each case, the SDCs of these features were concatenated with their original features and used as the discriminative features for the cognitive load classification experiment. The mel scale was chosen for this analysis since it is commonly used in speech-based cognitive load classification, and produced the highest performance for SDCs based on the spectral centroid features on both databases (Table 2). The resulting accuracies obtained are plotted in Fig. 7, together with those for MFCCs (including all DCT coefficients in each case).

The cognitive load classification accuracies shown in Fig. 7 indicate that the optimal number of filters is 6.

6.2. Effectiveness of different frequency scales

Mel scale filters have been commonly used to extract cepstral coefficients for speech recognition and speaker verification systems. This scale is a perceptually motivated scale devised through human perception experiments. However despite its popularity, it has been shown that the mel scale may not be the optimal scale for speech recognition (Shannon and Paliwal, 2003), and speaker identification systems (Lu and Dang, 2007). In this section the effects of using the mel, Bark, equivalent rectangular bandwidth (ERB), and hertz scales for extracting the *SCF* and *SCA* features are empirically analyzed through classification experiments. Like the mel scale, the ERB and Bark scales are also perceptually motivated scales where the frequency resolution is high in the low frequency region and low in high frequency as shown in Fig. 8. The mel scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another while Bark and ERB are the psychoacoustical scales whose bands are related to the critical bands of human auditory system. In terms of frequency resolution, the ERB provides the highest resolution in the frequency below 1 kHz while the mel provides the highest resolution in the frequency region above 4 kHz as shown in Fig. 8. These frequency scales have been used in speech recognition (Shannon and Paliwal,

Table 2
Classification accuracies (%) of SCF and SCA with different frequency scales.

System		mel scale		Bark scale		ERB scale		Hertz scale	
		Without dynamic	With dynamic	Without dynamic	With dynamic	Without dynamic	With dynamic	Without dynamic	With dynamic
Stroop test	SCF	49.3	82.0	61.5	84.3	59.1	82.8	40.0	71.3
	SCA	59.6	83.7	56.9	83.9	61.1	84.3	48.5	75.9
	Fusion SCF and SCA	60.9	87.2	62.8	84.6	60.3	86.5	48.1	77.6
Reading and comprehension	SCF	45.2	63.0	41.5	68.9	40.0	65.2	38.5	44.4
	SCA	46.7	61.5	47.4	64.4	43.7	63.7	42.2	47.4
	Fusion SCF and SCA	49.6	71.9	48.1	70.4	45.2	69.6	43.0	48.9

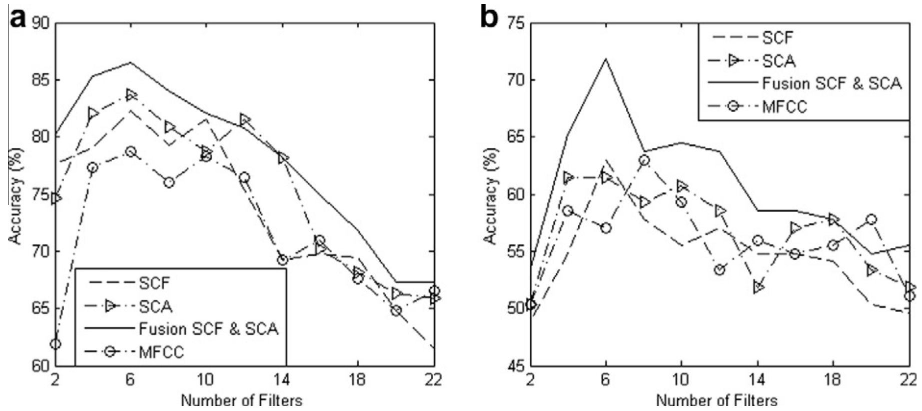

Fig. 7. Performance of SCF, SCA and MFCC with different numbers of subbands on (a) Stroop test corpus, and (b) Reading and Comprehension corpus.

2003) and stress detection (He et al., 2009). For the hertz scale, the filters are uniformly allocated along the bandwidth of speech. The choice of six filters is justified in Section 6.1.

The accuracies in Table 2 indicate that the spectral centroid features computed based on the three perceptual frequency scales, mel, Bark, and ERB, perform significantly better than those computed on the hertz frequency scale. This is most probably because compared with the hertz scale, the three perceptual scales provide a significantly higher resolution in the frequency region below 2 kHz (Fig. 8), which contains the most cognitive load specific information as suggested by Fig. 6. The use of all three perceptual frequency scales results in comparable classification accuracies. When SDCs are used to capture dynamic information, the mel frequency scale provides a marginal performance improvement compared with the Bark and the ERB scales for the CL classification system based on the fusion of the classification results of the systems using individual spectral centroid features.

Empirical frequency scales for each corpus based on the distribution of cognitive load information across different frequency bands were also investigated in our study. Specifically, the empirical scales for the two corpora were set up based on the spectral distribution of KL distance between CL models for SCA_m and SCF_m (Fig. 6). These curves were normalised to have the unit area below the curves prior to averaging them to obtain a single curve for each corpus. The empirical filter bank was then set up such that the six subbands divided the area under the overall KL distance curve into six regions whose areas are equal. The centre frequencies and the bandwidths of the six filters were then obtained as the centres and the widths of these subbands. This approach of filter allocation provides higher resolution in frequency ranges that contribute more towards discrimination between CL levels. The band allocations for the empirical scales are shown in Fig. 8. A similar approach has been used previously to design a 20-filter empirical frequency scale to capture the cepstrum coefficients for a CL classification system (Le et al., 2009). Performed on the Stroop test corpus, the SCF and SCA extracted using the empirical frequency scale combined with their SDC dynamic features provide accuracies of 85.9% and 83.3% and fusing the classification results of the two systems based on the individual features yields 87.8%. Performed on the Reading and Comprehension corpus, the corresponding results were 62.2%, 57.8% and 69.6% (fused). The performances of the empirical frequency scales are comparable to the perceptual frequency scales, probably because all these scales have similarly high resolution in the frequency region below 2 kHz.

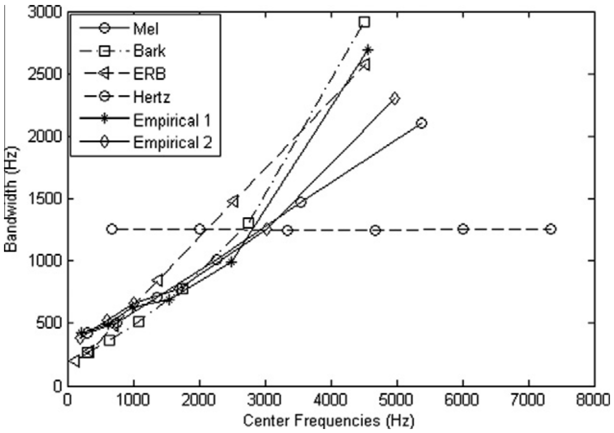

Fig. 8. Frequency allocation of filterbank centre frequencies and bandwidths. Empirical 1 and Empirical 2 are the empirical frequency scales used for Stroop test and Reading and Comprehension corpora respectively.

6.3. Combination of the two spectral centroid features

As expected, the fusion of the classification results at the score level of the SCF based system and SCA based system consistently improves the performance of the cognitive load classification systems. The results in Table 2 show that with six filters and various frequency scale configurations, when the SDCs were used, the fused system provided up to 24% and 27% reduction of the relative error rate compared with the systems based on individual features, for the Stroop test and Reading and Comprehension corpora respectively. The average reductions of relative error rate across the four frequency scales used are 18% and 14.1% for the Stroop test corpus and the Reading and Comprehension corpus. The effect of varying the weighting coefficients on the performance of the fused systems with six filters and various frequency scale configurations is highlighted in Fig. 9. As can be observed from this figure, in most cases, the value of weighting coefficient that maximises the performance of the system is around 0.5, suggesting that the contributions of SCF and SCA to the fused system are more or less equal.

Moreover, when the mel scale was utilised with a different number of filters, as shown in Fig. 7, the fused system provided performance improvements for almost all the cases. In particular, the fused system provided up to 29.5% and 26.9% of reduction in relative error rate for the Stroop test and the Reading and comprehension corpus.

Furthermore, although the performances of the systems based on individual spectral centroid feature are comparable with those of the systems based on the MFCCs, the combination of these two spectral centroid features at the score level consistently provides significantly better performance than those of the MFCCs systems as shown in Fig. 7. This is most likely because the combination of the two spectral centroid features capture the distribution of the spectrum in a subband more comprehensively than the MFCCs.

6.4. Performance comparison with the baseline system and state of the art system

It can be observed from Tables 1 and 2 that when static features are used, systems based on individual spectral centroid features extracted with six filters arranged in any perceptual frequency scale result in performances comparable to the baseline system. However, when the concatenation of the SDCs and the original feature is used as discriminative feature, systems that use spectral centroid features consistently provide better performances than the baseline system. Furthermore, the system based on fusing the classification results of the CL classification systems based on individual spectral centroid feature provides significantly higher performance when compared to the baseline system. In particular, the classification accuracies of the system using filters arrayed equally in the mel scale obtained when tested on the Stroop test corpus is 87.2%. Compared with the performance of the baseline system whose accuracy is 78.9%, this fused system provides a 39.3% relative reduction in error

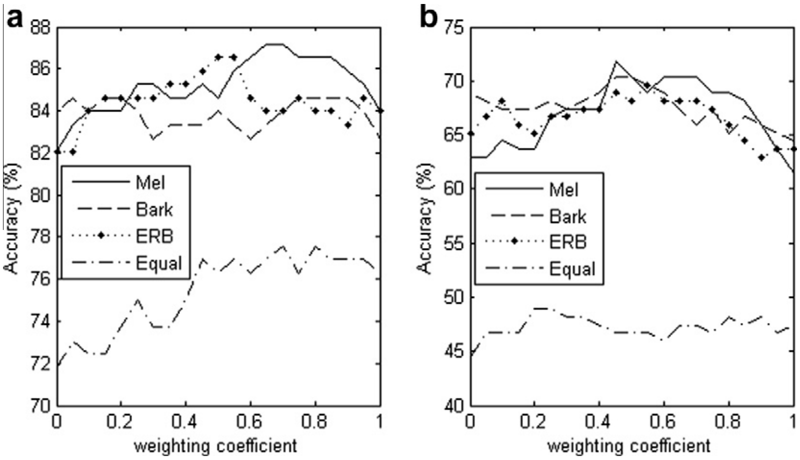

Fig. 9. Performance of the fused systems with various weighting coefficients (a) Stroop test corpus, and (b) Reading and Comprehension corpus.

Table 3
Cognitive load classification system performances (The entries in bold correspond to the highest classification accuracies).

System	Stroop test		Reading and Comprehension	
	Accuracy (%)	Relative improvement of accuracy (relative reduction of error rate) compared with baseline (%)	Accuracy (%)	Relative improvement of accuracy (relative reduction of error rate) compared with baseline (%)
MFCC + prosodic (baseline)	78.9	0 (0)	60.7	0 (0)
SCF	82.0	3.9 (14.7)	63.0	3.8 (5.9)
SCA	83.7	6.1 (22.7)	61.5	1.3 (2.0)
Fusion baseline and SCF	84.6	7.2 (27.0)	65.9	8.6 (13.2)
Fusion baseline and SCA	85.9	8.9 (33.2)	68.1	12.2 (18.8)
Fusion SCF and SCA	87.2	10.5 (39.3)	71.9	18.5 (28.5)
Fusion SCF and SCA and baseline	**88.5**	**10.8 (45.5)**	**72.6**	**19.6 (30.3)**

rate. Similarly, when tested on the Reading and Comprehension corpus, the classification accuracies of the fusion based system and the baseline system are 71.9% and 60.7%, providing a relative reduction in error rate of 28.5%.

Cognitive load classification systems based on the spectral centroid features, extracted using a bank of six mel scale filters (based on the results included in Sections 6.1 and 6.2), were also fused with the baseline system presented in Section 5.3, and the results are presented in Table 3. As expected, the system obtained by fusing SCF, SCA based systems to the baseline system outperforms all the other systems, for both corpora. Particularly, classification accuracy of the fused system performing on the Stroop test corpus is 88.5%. Compared with the performance of the baseline system whose accuracy is 78.9%, this fused system provides a 45.5% relative reduction in error rate. Similarly, when test on the Reading and Comprehension corpus, the classification accuracies of the fused system and the baseline system are 72.6% and 60.7%, a 30.3% relative reduction in error rate is obtained.

The previous state of the art system, tested on the Stroop test corpus involved fusing the scores of the baseline system (combination of MFCC, pitch, and intensity) with the scores of a glottal parameter feature based system (Yap et al., 2010a) produces an accuracy of 84.4%. The fused system presented in this paper has an accuracy of 88.5%, a 26.3% reduction in relative error rate. Similarly, the best system previously tested on the Reading and Comprehension corpus concatenated MFCC, pitch, and intensity and their SDCs and was evaluated in a speaker-dependent mode giving an accuracy of 71.1% (Yin et al., 2008). Our speaker-dependent system based on the fusion of the SCF based system and the SCA based system has an accuracy of 84.3%, a 45.7% reduction in relative error rate.

7. Discussion and conclusion

In this paper, we have investigated the use of spectral centroid features in cognitive load classification systems and include experimental results that consistently indicate that these spectral centroid features contain information that can be exploited by cognitive load classification systems. More specifically, we have shown that fusing the results of the cognitive load classification systems based on each of the two spectral centroid features consistently provides higher classification accuracy than that of the baseline system based on the combination of the MFCC, pitch and intensity.

We have identified the frequency regions that are relatively more important than other regions for the purpose of cognitive load classification when employing energy and spectral centroid based features. It is also significant that the distribution of cognitive load specific information is consistent for both corpora tested, even though they have very different characteristics.

The accuracies of cognitive load classification systems when tested on the Reading and Comprehension corpus are consistently lower than those obtained from the Stroop test corpus. This is to be expected as the Reading and Comprehension corpus was recorded in a much less controlled manner than the Stroop test corpus. Although the Reading and Comprehension is still a corpus collected in a laboratory, it is closer to a realistic as it contains continuous and spontaneous speech. The relatively high accuracy obtained by cognitive load classification systems based on spectral centroid features when tested on this corpus is therefore promising, although there is clearly still room for further improvement.

Future work will include validating these results on other cognitive load speech corpora.

References

Berthold, A., Jameson, A., 1999. Interpreting symptoms of cognitive load in speech input. In: Proc. Internat. Conf. on User Modeling, 1999, pp. 235–244.

Boril, H., Sadjadi, O., Kleinschmidt, T., Hansen, J.H.L., 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In: Proc. Interspeech, Makuhari, Chiba, Japan, 2010, pp. 502–505.

Fernandez, R., Picard, R.W., 2003. Modeling drivers' speech under stress. Speech Commun. 40, 145–159.

Gajic, B., Paliwal, K.K., 2006. Robust speech recognition in noisy environments based on subband spectral centroid histogram. IEEE Trans. Speech Audio Lang. Process. 14, 600–608.

Gerven, P.W.M.V., Pass, F., Merrienboer, J.J.G.V., 2004. Memory load and the cognitive pupillary response in aging. Psychophysiology 41, 167–174.

Goldberger, J., Aronowitz, H., 2005. A distance measure between GMMs based on the unscented transform and its application to speaker recognition. In: Proc. Interspeech, Lisboa, Portugal, 2005, pp. 1985–1988.

Griffin, G., Williams, C., 1987. The effects of different levels of task complexity on three vocal measures. Aviation, Space, Environ. Med. 58, 1165.

He, L., Lech, M., Maddage, N.C., Allen, N., 2009. Stress detection using speech spectrograms and sigma-pi neuron units. In: Proc. Fifth Internat. Conf. on Natural Computation, Tianjin, 2009, pp. 260–264.

Hosseinzadeh, D., Krishnan, S., 2008. On the use of complementary spectral features for speaker recognition. EURASIP J. Adv. Signal Process., 1–10.

Khawaja, M.A., Ruiz, N., Cheng, F., 2007. Potential speech features for cognitive measurement. In: Proc. 19th Australian Conf. on Computer–Human Interaction: Entertaining User Interfaces, Adelaide, Australia, 2007, pp. 57–60.

Kleinschmidt, M., 2002. Methods for capturing spectro-temporal modulations in automatic speech recognition. Acoust. United Acta Acoust. 88, 416–422.

Kua, J.M.K., Thiruvaran, T., Nosratighods, M., Ambikairajah, E., Epps, J., 2010. Investigation of spectral centroid magnitude and frequency for speaker recognition. In: Proc. Odyssey, The Speaker and Language Recognition Workshop, Brno, Czech Republic, 2010, pp. 34–39.

Le, P.N., Ambikairajah, E., Choi, E.H.C., Epps, J., 2009. A non-uniform subband approach to speech-based cognitive load classification. In: Proc. Seventh Internat. Conf. on Information, Communications and Signal Processing, Fisherman's Wharf, Macau, 2009, pp. 1–5.

Le, P.N., Epps, J., Choi, E.H.C., Ambikairajah, E., 2010. A study of voice source and vocal tract filter based features in cognitive load classification. In: Proc. 20th Internat. Conf. on Pattern Recognition, Istanbul Turkey, 2010, pp. 4516–4519.

Lively, S., Pisoni, D., Van Summers, W., Bernacki, R., 1993. Effects of cognitive workload on speech production: acoustic analyses and perceptual consequences. J. Acoust. Soc. Amer. 93, 2962–2973.

Lu, X., Dang, J., 2007. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Commun., 312–322.

Mendoza, E., Carballo, G., 1998. Acoustic analysis of induced vocal stress by means of cognitive workload tasks. J. Voice 12, 263–273.

Metametrics, The Lexile Framework for Reading, 2007.

Muller, C., Grossmann-Hutter, B., Jameson, A., Jummer, R., Wittig, F., 2001. Recognizing time pressure and cognitive load on the basis of speech: an experimental study. Lecture Notes Comput. Sci., 24–33.

Paas, F., Tuovinen, J.E., Tabbers, H., Gerven, P.W.M.V., 2003. Cognitive load measurement as a means to advance cognitive load theory. Educ. Psychol. 38, 63–71.

Paliwal, K.K., 1998. Spectral subband centroid features for speech recognition. In: Proc. ICASSP, Seattle, WA, USA, 1998, pp. 617–620.

Pass, F.G.W.C., Merrienboer, J.J.G.V., 1994. Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive-load approach. J. Educ. Psychol. 86, 122–133.

Pelecanos, J., Sridharan, S., 2001. Feature warping for robust speaker verification. ODYSSEY, 213–218.

Peterson, G.E., Barney, H.L., 1952. Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175–184.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10, 19–41.

Scherer, K.R., Grandjean, D., Johnstone, T., Klasmeyer, G., Bänziger, T., 2002. Acoustic correlates of task load and stress. In: Proc. ICSLP, Denver, Colorado, USA, 2002, pp. 2017–2020.

Shannon, B.J., Paliwal, K.K., 2003. A comparative study of filter bank spacing for speech recognition. In: Proc. Microelectronic Engineering Research Conference, Brisbane, Australia, 2003, pp. 1–3.

Shriberg, E., Bear, J., Dowding, J., 1992. Automatic detection and correction of repairs in human–computer dialog. In: Proc. Fifth DARPA Speech and Natural Language Workshop, Morgan Kaufmann, San Mateo, pp. 419–424.

Steeneken, H.J.M., Hansen, J.H.L., 1999. Speech under stress conditions: overview of the effect on speech production and on system performance. In: Proc. ICASSP, Phoenix, AZ, USA, 1999, pp. 2079–2082.

Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier Science B.V., Amsterdam, pp. 495–518.

Thiruvaran, T., Ambikairajah, E., Epps, J., 2006. Speaker identification using FM features. In: Proc. 11th Australia Internat. Conf. on Speech Science and Technology, Canberra, Australia, 2006, pp. 148–152.

Yap, T.F., Ambikairajah, E., Choi, E., Chen, F., 2009. Phase-based features for cognitive load measurement system. In: Proc. ICASSP, Taipei, Taiwan, 2009, pp. 4825–4828.

Yap, T.F., Ambikairajah, E., Epps, J., Choi, E.H.C., 2010. Cognitive load classification using formant features. In: Proc. ISSPA, Kuala Lumpur, Malaysia, 2010, pp. 221–224.

Yap, T.F., Epps, J., Choi, E.H.C., Ambikairajah, E., 2010a. Glottal features for speech-based cognitive load classification. In: Proc. ICASSP, Dallas, Texas, USA, 2010a, pp. 5234–5237.

Yap, T.F., Epps, J., Ambikairajah, E., Choi, E.H.C., 2010b. An investigation of formant frequencies for cognitive load classification. In: Proc. InterSpeech, Makuhari, Chiba, Japan, 2010b, pp. 2022–2025.

Yin, B., Ruiz, N., Chen, F., Khawaja, M.A., 2007. Automatic cognitive load detection from speech features. In: Proc. CHISIG, Adelaide, Australia, 2007, pp. 249–255.

Yin, B., Chen, F., Ruiz, N., Ambikairajah, E., 2008. Speech-based cognitive load monitoring system. In: Proc. ICASSP, Las Vegas, Nevada, USA, 2008, pp. 2041–2044.

Please cite this article in press as: Le, P.N. et al., Investigation of spectral centroid features for cognitive load classification, Speech Comm. (2011), doi:10.1016/j.specom.2011.01.005