A Study of Voice Source and Vocal Tract Filter Based Features in Cognitive Load Classification

Phu Ngoc Le[1,2], Julien Epps[1,2], Eric H. C. Choi[2], Eliathamby Ambikairajah[1,2]
[1]*School of Electrical Engineering and Telecommunications*
The University of New South Wales, UNSW Sydney, NSW 2052, Australia.
[2]*ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia.*
phule@unsw.edu.au, j.epps@unsw.edu.au, eric.choi@nicta.com.au, ambi@ee.unsw.edu.au

Abstract

Speech has been recognized as an attractive method for the measurement of cognitive load. Previous approaches have used mel frequency cepstral coefficients (MFCCs) as discriminative features to classify cognitive load. The MFCCs contain information from both the voice source and the vocal tract, so that the individual contributions of each to cognitive load variation are unclear. This paper aims to extract speech features related to either the voice source or the vocal tract and use them to discriminate between cognitive load levels in order to identify the individual contribution of each for cognitive load measurement. Voice source-related features are then used to improve the performance of current cognitive load classification systems, using adapted Gaussian mixture models. Our experimental result shows that the use of voice source feature could yield around 12% reduction in relative error rate compared with the baseline system based on MFCCs, intensity, and pitch contour.

1. Introduction

Cognitive load (CL) refers to the amount of mental demand imposed on the human cognitive capacity when performing a particular task [1]. Measuring cognitive load, or classifying along an ordinal CL scale, is important in designing an optimal interaction approach between humans and computing systems in order to produce the highest task performance. Speech has been recognized as a good approach for CL measurement due to its non-intrusive and inexpensive attributes as a system input. Several types of speech features, e.g., mel-frequency cepstral coefficient (MFCC), pitch, intensity, frequency modulation, and group delay, have been proposed to classify cognitive load, of which the most successful individual feature is MFCC [2, 3]. In linear acoustic theory, the speech production process is described in terms of the voice source excitation and vocal tract filter (Fig. 1), and information from both of these components is present in MFCCs. To date, the cognitive load discrimination due to the voice source and the vocal tract filter has not been established.

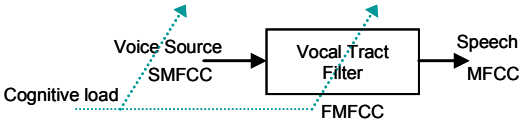

Figure 1. The speech production model

In our recent study [4], we analyzed the cognitive load discrimination due to features extracted in different frequency bands of speech, and showed that features derived from bands from 0 to 800 Hz significantly outperform those from the other bands. In this range the voice source spectrum contains high energy, and we hypothesize that features from the voice source are important for cognitive load classification.

When voiced speech is generated, the voice source can be characterized by the glottal waveform, and glottal waveform-derived features have recently been reported in emotion, deceptive/non-deceptive, and clinical depression classification systems [5].

In this study, we extract speech features related specifically to the voice source and the vocal tract filter and evaluate their performance in a UBM-GMM cognitive load classification system [3], both individually and combined using fusion.

2. Cognitive load classification system

2.1. Source and filter separation

Based on the linear acoustic model of Fig. 1, the Iterative Adaptive Inverse Filtering (IAIF) algorithm was proposed to estimate the glottal waveform of

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1097

speech signal by filtering the original speech signal using an inverse model of the vocal tract filter, modeled as an all-pole system. The detailed description of the IAIF algorithm can be found in [6]. In our study, IAIF is used to estimate both the glottal waveform and all-pole model parameters for the vocal tract filter in isolation. The integration of IAIF into our feature extraction and CL classification system is illustrated in Fig. 2.

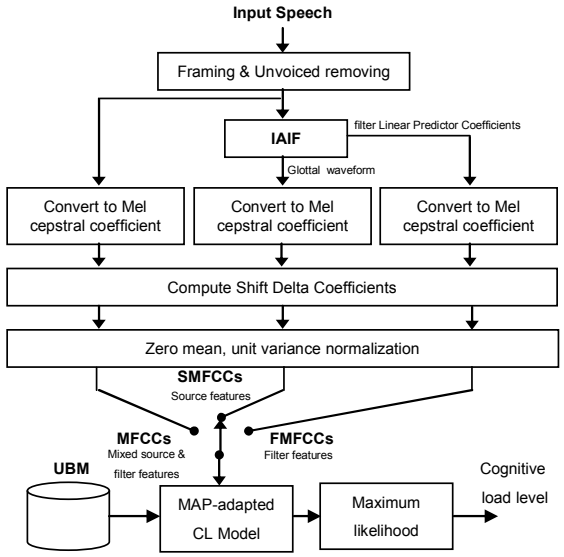

Figure 2. Cognitive load classification system

2.2. Source and filter feature extraction

The input speech is segmented into 25 ms length, 15 ms overlapped frames. The unvoiced frames are detected and removed using voiced activity detection (VAD), which detects unvoiced frames based on the result of pitch estimation. Three sets of speech features are then extracted from each frame of voiced speech:

Mel Frequency Cepstrum Coefficients (MFCCs); 12 MFCCs are extracted from 29 mel-scale filters in the 0-8 kHz range (excluding the zero order MFCC).

Source Mel Frequency Cepstrum Coefficients (SMFCCs); The computation of SMFCCs follows exactly the same steps as for computing MFCCs, except that the input is the glottal waveform output by IAIF rather than the speech signal.

Filter Mel Frequency Cepstrum Coefficients (FMFCCs); The spectral envelope of the vocal tract filter is evaluated from the linear predictor coefficients, obtained from the implementation of the IAIF. Twelve *FMFCCs* are then obtained as the first 12 output coefficients (excluding the zero order) of the discrete cosine transform of the logarithm of 29 mel-scale filter energies, derived from the magnitude response of the linear predictive filter.

The MFCC parameterization was chosen to describe the glottal waveform and the vocal tract filter in our study because the direct comparison of accuracy between SMFCCs, FMFCCs, and (speech) MFCCs may lead to a more straightforward comparison between the intrinsic cognitive load discrimination power due to the source and filter features.

2.3 Dynamic features

All feature vectors mentioned above are static features, i.e., each feature element only represents the feature data at exactly the point of calculation. In this study, the SDC technique is used to capture the temporal evolution of MFCC, SMFCC and FMFCC feature vectors, as it has been used with success in CL classification previously [3], and allows control over the temporal duration of the dynamic features.

Mathematically, SDC feature vectors are computed from the original static feature vector as

$$\mathbf{F}_{SDC}(t)=conc(\mathbf{c}(t+iP+D)-\mathbf{c}(t+iP-D))_{i=0\to k}, \qquad (1)$$

where t is the time (frame number) when SDC is computed; ***conc*(.)** is the concatenation operation; **c** is the original static feature vector; and D, P and k are parameters of SDC. In our study, $D = 3$, $P = 3$, and $k \in [0, 7]$.

Feature warping is utilized in this paper to map the distribution of the feature vectors over each utterance to a distribution with zero mean, unit variance.

2.4 Classification

A UBM-GMM based classifier [7] is used as the back-end of the classification system in this paper. Each cognitive load level is modeled as a Gaussian mixture model (GMM). A GMM UBM (Universal Background Model) is trained based on the normal speech of all speakers for a set of reading tasks. The speech features of each cognitive load level are then used to adapt this UBM to obtain CL models using the maximum a posteriori (MAP) technique [7], illustrated in Fig. 3. The sufficient statistics from the training data of a cognitive load level are used to update the UBM sufficient statistics for a particular mixture, creating adapted parameters on a per-mixture basis. For instance, the adapted mean of the ith mixture ($\hat{\mu}_i$) is obtained from the ith UBM mixture mean (μ_i):

$$\hat{\mu}_i = \alpha_i^m E_i(x) + (1-\alpha_i^m)\mu_i, \qquad (2)$$

where $E_i(x)$ is the mean parameter computed from the probabilistic alignment process of the training vectors into the ith UBM mixture, and α_i^m are the adaptation coefficients. α_i^m are data-dependent, so mixtures with

a high count of data from the cognitive load-specific training rely more on the new sufficient statistics ($\alpha_i^m \to 1$, c.f. long arrows in Fig. 3) and mixtures with low counts of data rely more on the UBM sufficient statistics ($\alpha_i^m \to 0$, c.f. short arrow in Fig. 3).

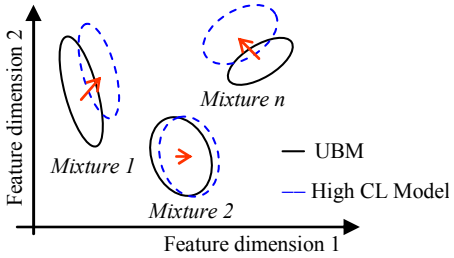

Figure 3. Conceptual diagram showing adaptation of a high cognitive load model from a UBM

In the testing phase, features extracted from test speech are used to determine the likelihood scores from the GMMs. The classification result is then obtained as the best matched model (maximum log-likelihood score *LL*). The number of mixtures of the UBM and all GMMs in our study is 256. In order to improve the reliability of the results, all experiments in this paper were performed in a leave-one-out cross-validation fashion where data from speakers appearing in the test data were not present in the training data.

2.5 Score-level fusion

A linear search fusion technique was employed to combine the log-likelihood scores of individual classifiers based on different features. Fusion allows us to investigate the complementary behavior of SMFCC and FMFCC in characterizing the cognitive load variation. The fused log-likelihood score was obtained from the linear combination of log-likelihood scores LL_1 and LL_2 from two individual systems as follows:

$$LL_{fused} = \alpha LL_1 + (1-\alpha) LL_2 \qquad (3)$$

where LL_{fused} is the fused log-likelihood score, and $0 \le \alpha \le 1$ is the weighting coefficient empirically chosen to optimize the performance of the system.

3. Evaluation

3.1 Database

All experiments in this paper were performed on the Stroop test corpus, containing speech elicited under three levels of cognitive load (low, medium, and high) from 15 native English speakers (8 females and 7 males) performing the Stroop test [3]. The low CL task required subjects to read the color name of words written in black or congruent font color (font color is same as color word). The medium CL task required them to read the font color of words in incongruent color (font color is different with color word). The high CL task was the same as for medium CL, except that a time constraint was added. The database contains approximately 60 seconds of speech per cognitive load level per speaker. We also used story reading speech (low CL, 90 sec per subject) collected from the same subjects to train the UBM.

3.2 Results: Static features

SMFCCs, FMFCCs and MFCCs were individually used to classify CL, in order to evaluate the CL discrimination of speech features related to the voice source and vocal tract filter. The classifier results from systems based on different feature sets were then fused in pairs to investigate their complementary properties.

It is evident from Table 1 that significant cognitive load information is captured by both SMFCCs and FMFCCs. This suggests that cognitive load variation is characterized by both the voice source and vocal tract filter components of the human speech production model. However, it is interesting to notice that the use of SMFCCs yields higher accuracy than that of FMFCCs alone, suggesting for MFCC features, the voice source is marginally more important than vocal tract filter in terms of cognitive load variation.

Table 1. Cognitive load classification accuracy for static source and filter feature vectors

System	Accuracy (%)
SMFCCs	45
FMFCCs	43
MFCCs	50.4
Fusion SMFCCs & FMFCCs	49.4
Fusion SMFCCs & MFCCs	**54.5**
Fusion FMFCCs & MFCCs	50.6

The use of MFCCs produced higher accuracy than SMFCCs or FMFCCs alone, probably because MFCCs capture information related to both voice source and vocal tract filter. This result is confirmed by the expected result that the fusion of scores based on SMFCC and FMFCC features produces approximately the same accuracy as that for MFCCs alone. The fusion between SMFCC and MFCC based systems admits a further improvement of classification accuracy over MFCCs, which seems to suggest that MFCCs do not completely capture the discriminatory information relating to cognitive load from the voice source.

3.3 Results: Dynamic features

Dynamic features, based on SMFCCs, FMFCCs, and MFCCs obtained using the SDC technique with various values of *k*, were used to classify cognitive

load in order to evaluate the cognitive load discrimination ability of temporal information.

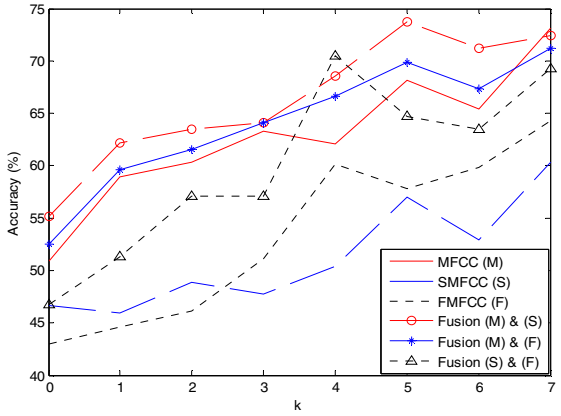

Figure 4. Cognitive load classification accuracy (%) using SDC feature vectors with various value of *k*

It can be seen from Fig. 4 that SDC significantly improves the performance of all classification systems, as in [3]. Increasing k convincingly improves classification accuracy, as more temporal information is captured when k increases, but perhaps surprisingly so given the high feature dimension of SDC relative to the small database used herein. For all values of k, among systems based on individual feature sets, the MFCC based system always yielded the highest classification accuracy which again confirms that MFCC captures CL discriminative information related to both voice source and vocal tract filter. Fusion between SMFCC and MFCC based systems generally yielded an increase in accuracy over any individual system, although the benefit over MFCCs declines as the overall accuracy increases. Accuracy improvements for the source and filter features when more temporal information is included may be smaller than those of the MFCCs due to the differing effects of the IAIF algorithm from one frame to the next. Our further experiments with $k>7$ show that the accuracy of the classification systems saturates at around $k=7$.

3.4 Results: State of the art baseline

To confirm the effectiveness of the voice source related features in discriminating cognitive load, the classification result of our SMFCC based system, employing SDC with k=7 was fused with that of the baseline system developed in [3]. The accuracy of the baselines system is 78.1% where the discriminatory feature vectors were SDCs (k=7) computed on a combination of pitch, intensity, and MFCCs. Fusion of this baseline with the SMFCC, k=7 system reported herein produced an accuracy of 80.8%, again suggesting that voice source related features can be used to improve the CL classification accuracy.

5. Conclusion

This paper has presented an investigation of the cognitive load discrimination ability due to voice source and vocal tract filter related information individually. The results of our study have partly validated the hypothesis that voice source related features are important in discriminating cognitive load level, as suggested by [4]. Incorporating more information from the voice source therefore has potential to improve the performance of current cognitive load classification systems, many of which exploit discriminative information mainly from the vocal tract filter. Future work will investigate features derived principally from the low-frequency components of the voice source.

6. References

[1] F. Paas, J. E. Tuovinen, H. Tabbers, and P. W. M. V. Gerven, "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," *Education Psychologist,* vol. 38, pp. 63-71, 2003.

[2] T. F. Yap, E. Ambikairajah, E. Choi, and F. Chen, "Phase-based features for cognitive load measurement system," *in Proc. of ICASSP,* pp. 4825-4828, 2009.

[3] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," *in Proc. of ICASSP,* pp. 2041-2044, 2008.

[4] P. N. Le, E. Ambikairajah, E. H. C. Choi, and J. Epps, "A Non-Uniform Subband Approach to Speech-Based Cognitive Load Classification," *the 7th ICICS, Dec. 2009, Macau*.

[5] J. F. Torres, E. Moore II, and E. Bryant, "A study of glottal waveform features for deceptive speech classification," *in Proc. of ICASSP,* pp. 4489-4492, 2008.

[6] P. Alku, "Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering," *In Proc. of EuroSpeech,* pp. 1081-1084, 1991.

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *in Proceedings of Digital Signal Processing,* vol. 10, pp. 19-41, 2000.