Acta Cybernetica **25** (2021) 223–232.

Using the Fisher Vector Approach for Cold Identification*

José Vicente Egas-López[a] and Gábor Gosztolya[b]

Abstract

In this paper, we present a computational paralinguistic method for assessing whether a person has an upper respiratory tract infection (i.e. cold) using their speech. Having a system that can accurately assess a cold can be helpful for predicting its propagation. For this purpose, we utilize Mel-frequency Cepstral Coefficients (MFCC) as audio-signal representations, extracted from the utterances, which allowed us to fit a generative Gaussian Mixture Model (GMM) that serves to produce an encoding based on the Fisher Vector (FV) approach. Here, we use the URTIC dataset provided by the organizers of the ComParE Challenge 2017 of the Interspeech Conference. The classification is done by a linear kernel Support Vector Machines (SVM). Owing to the high imbalance of classes on the training dataset, we opt for undersampling the majority class, that is, to reduce the number of samples to those of the minority class. We find that applying Power Normalization (PN) and Principal Component Analysis (PCA) on the Fisher Vector features is an effective strategy for the classification performance. We get a better performance than that of the Bag-of-Audio-Words approach reported in the paper of the challenge.

Keywords: computational paralinguistics, speech processing, machine learning, fisher vector

1 Introduction

Upper respiratory tract infection (URTI) is an infectious process for any of the components of the upper airway. E.g., the common cold, a sinus infection, amongst

*This study was supported by the Hungarian Artificial Intelligence National Laboratory, by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413, and by the grant NKFIH-1279-2/2020 of the Hungarian Ministry of Innovation and Technology. Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Programme ÚNKP-20-5.

[a]Institute of Informatics, University of Szeged, Hungary, E-mail: egasj@inf.u-szeged.hu, ORCID: 0000-0002-5622-9192

[b]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary, E-mail: ggabor@inf.u-szeged.hu, ORCID: 0000-0002-2864-6466

DOI: 10.14232/actacyb.287868

others. Being able to automatically assess whether a subject has a cold may be relevant when trying to prevent the spread of it by predicting its patterns of propagation. The area of computational paralinguistics differs from Automatic Speech Recognition (ASR), which focuses on the actual *content* of the speech of an audio signal. Here, computational paralinguistics may provide the necessary tools for determining the *way* the speech is spoken. Various studies have offered promising results in this area: diagnosing neuro-degenerative diseases using the speech of the patients [5, 6, 7], the classification of crying sounds and heart beats [10], estimating the sincerity of apologies [9], determining the depression of a subject [4]. In this study, we focus on finding specific voice patterns latent in the speech of subjects having a *cold*.

Previous studies applied various approaches for classifying *cold* subjects using the same corpus. For example, Gosztolya et al. employed Deep Neural Networks for feature extraction for this purpose [8]. Huckvale and Beke utilized four types of voice features for studying changes in health [11]. Furthermore, Kaya et al. [14] introduced the application of a weighting scheme on instances of the corpus, making use of a Weighted Kernel Extreme Learning Machine in order to handle the imbalanced data that comprises the URTIC corpus. As any other computational paralinguistic task, assessing a cold from the speech is a challenging issue. Finding out the latent patterns that could characterize or represent a cold speech does not only depend on the feature extraction phase but in the data itself too. This may be attributed to different perspectives: limited amount of data, data imbalance, quality of the recordings.

In this study, we exploit the Upper Respiratory Tract Infection Corpus (URTIC that was the dataset of one of the Sub-Challenges in the ComParE Challenge from Interspeech 2017) [21]. In the feature extraction phase, we selected frame-level features. Namely, we utilize Mel-frequency Cepstral Coefficients (MFCC) as audio-signal representations, extracted from the utterances. This allowed us to fit a generative Gaussian Mixture Model (GMM) that can produce an encoding based on the Fisher Vector (FV) approach. That is, the computation of low-level patch descriptors together with their deviations from the GMM gives us an encoding (i.e. feature) called the Fisher Vector.

Unweighted Average Recall (UAR) scoring was used to measure the performance of the model since it is the de facto standard metric for these kinds of challenges [18]. To the best of our knowledge, this is the first study that focuses on making use of a FV representation in order to detect a cold.

Furthermore, we find that applying Power Normalization (PN) and Principal Component Analysis (PCA) on the Fisher Vector features is an effective strategy for the classification performance. In the next part of our study, we employ a late-fusion of the ComParE Bag-of-Audio-Words (BoAW) features with the Fisher Vector representations. Mentioned fusion technique contributes to the classification performance.

Table 1: Upper Respiratory Tract Infection Corpus (URTIC).

Class	**Train**	**Development**	**Test**	**Total**
Cold	970	1011	895	2876
Not-Cold	8535	8585	8656	25,776
Total	9505	9596	9551	28,652

2 Data

The entire dataset consists of 382 male speakers, 248 female speakers, with a mean age of 29.5 years; and a sampling rate of 44.1kHz downsampled to 16kHz. For the Sub-Challenge, the corpus was provided by the Institute of Safety Technology, University of Wuppertal, Germany. The following tasks were performed by the participants: they had to read short stories (e.g. the well-known story in the field of phonetics *The North Wind and the Sun*, to produce voice commands (such as numbers from 1 to 40), and to narrate spontaneous speech (i.e. tell something about their last weekend or their best vacation). Note that the number of tasks varied for each speaker. The recordings were split into 28,652 chunks of 3 to 10 seconds in length. Specifically, the division of the chunks was carried out in a speaker-independent manner, each partition (i.e. train, development, and test) having 210 speakers. The training and development sets are both comprised by 37 subjects having a cold and 173 subjects not having a cold. The reader may see more details in [22]. The number of samples and classes for each dataset is described in Table 1.

3 Methodology

Figure 1 shows the methodology employed in this study: (1) Frame-level features (MFCC) were extracted from the utterances; (2) A Gaussian Mixture Model (GMM) is trained utilizing the MFCC representations; (3) Fisher Vector features are extracted using the trained GMM; and (4) SVM performs the classification task.

3.1 Frame-level feature extraction

The features we employed were the well-known MFCCs with a dimension of 13, along with their first and second order derivatives, frame-length and frame-shift of 25 ms and 10 ms, respectively. We used the Kaldi Speech Recognition Toolkit [17] for this task.

3.2 Fisher Vector (FV)

The FV approach is an image representation that pools local image descriptors [19]. It was originally intended for image classification but here we exploit its ability to generate a complete representation of the samples which are later characterized by their deviation from a generative GMM. The samples can be thought of as local patch descriptors of an image. In our case, they are the frame-level features of an audio signal. FV is an improved version of the general case called the Fisher Kernel (FK) [12], which measures the similarity of two objects from a parametric generative model of the data. The FK will be explained more in detail in the next section. FV basically assigns a local descriptor to elements in a visual dictionary. This approach stores visual word occurrences and takes into account the difference between dictionary elements and pooled local features, it stores their statistics as well.

3.2.1 Fisher Kernel (FK)

It seeks to measure the similarity of two objects from a parametric generative model of the data (X) which is defined as the gradient of the log-likelihood of X [12]:

$$G_\lambda^X = \bigtriangledown_\lambda \log \upsilon_\lambda(X), \tag{1}$$

where $X = \{x_t, t = 1, \ldots, T\}$ is a sample of T observations $x_t \in \mathcal{X}$, υ represents a probability density function that models the generative process of the elements in $\mathcal{X}$ and $\lambda = [\,\lambda_1, \ldots, \lambda_M]\,' \in R^M$ stands for the parameter vector υ_λ [19]. Thus, such a gradient describes the way the parameter υ_λ should be changed in order to best fit the data X. A way to measure the similarity between two points X and Y by means of the FK can be expressed as follows [12]:

$$K_{FK}(X, Y) = G_\lambda^{X\prime} F_\lambda^{-1} G_\lambda^Y. \tag{2}$$

Eq. (3) shows how the Cholesky decomposition $F_\lambda^{-1} = L'_\lambda L_\lambda$ can be utilized to rewrite the Eq. (2) in terms of the dot product:

$$K_{FK}(X, Y) = \mathscr{G}_\lambda^{X\prime} \mathscr{G}_\lambda^Y, \tag{3}$$

where

$$\mathscr{G}_\lambda^X = L_\lambda G_\lambda^X = L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X). \tag{4}$$

Such a normalized gradient vector is the so-called *Fisher Vector* of X [19]. Both the FV $\mathscr{G}_\lambda^X$ and the gradient vector G_λ^X have the same dimension.

3.2.2 Fisher Vector for audio-signals

Let $X = \{X_t, t = 1 \ldots T\}$ be the set of D-dimensional local SIFT descriptors extracted from an image and let the assumption of independent samples hold, then Eq. (4) becomes:

$$\mathscr{G}_\lambda^X = \sum_{t=1}^{T} L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X_t). \tag{5}$$

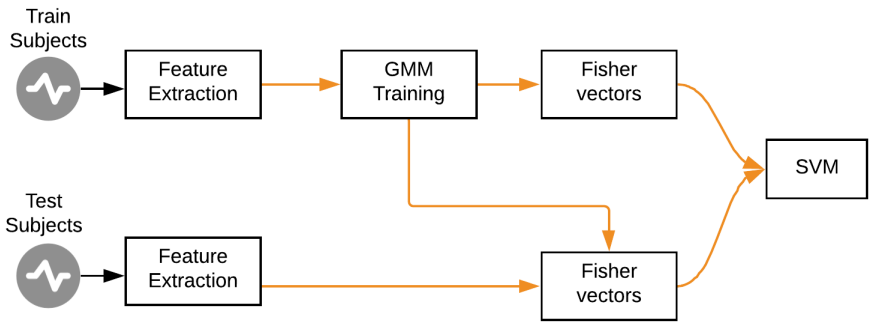

Figure 1: The methodology applied in this study.

The assumption of independence permits the FV to become a sum of normalized gradients statistics $L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(x_t)$ calculated for each SIFT descriptor:

$$X_t \to \varphi_{FK}(X_t) = L_\lambda \bigtriangledown_\lambda \log \upsilon_\lambda(X_t), \tag{6}$$

which describes an operation that can be thought of as a higher dimensional space embedding of the local descriptors X_t.

Hence, the FV approach extracts low-level local patch descriptors from the audio-signals' spectrogram. Then, with the use of a GMM with diagonal covariances we can model the distribution of the extracted features. The log-likelihood gradients of the features modeled by the parameters of such GMM are encoded through the FV [19]. This type of encoding stores the mean and covariance *deviation* vectors of the components k that form the GMM together with the elements of the local feature descriptors. The image is represented by the concatenation of all the mean and the covariance vectors that gives a final vector of length $(2D + 1)N$, for N quantization cells and D dimensional descriptors [16, 19].

The FV approach can be compared with the traditional encoding method: BoV, and with a first order encoding method like Vector of Locally Aggregated Descriptors (VLAD) [1]. In practice, BoV and VLAD are outperformed by FV due to its second order encoding property of storing additional statistics between codewords and local feature descriptors [23].

The FV representation, regardless of the number of local features (i.e. SIFT), or in our case, frame-level features (MFCCs), extracts a *fixed-sized* feature representation from each image (i.e. from each MFCC representation). Here, we use FV features to encode MFCC features extracted from audio-signals of HC and PD subjects. FV allows us to give a complete representation of the sample set by encoding the count of occurrences and higher-order statistics associated with its distribution.

3.3 Classification

Support Vector Machines (SVM) is the classification algorithm used to assess the recordings, it is typically the standard choice for paralinguistics tasks. Moreover, this algorithm can achieve good performances when used with FV [19, 24]. As for the evaluation metric, Unweighted Average Recall (UAR) is the proper way to measure the performance of these kinds of tasks; principally because it is commonly used when there is the need to handle class imbalance situations. Furthermore, this metric has been utilized since the very first ComParE Challenge (see [20] for more details about the UAR evaluation metric).

4 Experimental Setup

The training dataset consists of 9505 utterances, where 8535 (89.8%) are labeled as *healthy* (not-cold) and the rest, 970 (10.2%), are labeled as *cold*. Likewise, the development dataset comprises 1011 *cold* and 8585 *not-cold* labels, which are 10.53% and 89.47%, respectively. Such a high class imbalance is more likely to diminish the performance of the SVM classifier. To overcome this, we used random undersampling which reduces the number of samples associated with all classes to that of the minority class, i.e. *cold*. We relied on imbalanced-learn [15], which is a Python package offering several resampling methods used in datasets that have a between-class imbalance. In our first experiments we reduced the dimensions of the features via Principal Component Analysis (PCA) [13], keeping a variance of 0.95. Chatfield et al. demonstrate that applying PCA before classification enhances the discrimination task with FV and reduces the memory consumption as well [3].

Moreover, the features (Fisher Vectors) were normalized with Power Normalization (PN) and *l2-Normalization*. Power Normalization was found to be helpful for FVs representations [19] as it reduced the impact of the features that become more sparse as the number of Gaussian components increases. In the following experiments, we applied these normalization techniques before reducing the dimensions using PCA. Likewise, we found that *l2-Norm.* helped to alleviate the effect of having different utterances with distinct amounts of background information projected into the extracted features, which attempts to improve the prediction performance.

The GMM used in our experiments to compute the FVs was set to operate with a varying number of components: G_c ranged from 2 to 128. The construction of the Fisher Vector representations was made with the help of a Python-wrapped version of the VLFeat library [25]. As stated before, the classification was done using a Support Vector Machines algorithm. We employed the libSVM implementation [2] with a linear kernel and, as suggested in [12], the C complexity parameter was set in the range 10^{-5}, ..., 10^{1}. In order to search for the best complexity value (C) of the SVM, Stratified Group k-fold Cross Validation (CV) was applied over the training and development sets. This type of CV allowed us to avoid having the same speaker in more than one specific fold, while simultaneously preserving the percentage of samples for each class within each fold.

Table 2: UAR scores obtained when SVM classified the data using Fisher Vectors.

Features	**GMM size**	**Performance (%)**	
		Cross-Val	**Test**
ComParE (BoAW-baseline)	-	64.54%	67.30%
Fisher Vectors	64	63.98%	66.12%
Fisher Vectors + PCA	64	64.72%	67.65%
Fisher Vectors + PN + PCA	64	64.92%	67.81%
ComParE + Fisher Vectors (+PN+PCA)	-	63.01%	70.17%

Finally, we performed *late-fusion* of the best configurations. Namely, the class-wise posterior estimates generated by the SVM algorithm could provide a simple way of classifier combination by taking the mean of two or more posterior vectors.

5 Results

As shown in Table 2, for the baseline we utilized the ComParE functionals (i.e. Bag-of-Audio-Words features) that were originally presented and described in [21]. According to the results outlined in Table 2, these representations achieved an UAR score of 67.3% on the test set. This score was slightly outperformed by two of our configurations: when PCA was applied (67.65%), and when PN was applied along with PCA (67.81%). Table 1 shows the results obtained when using Fisher Vectors with their complete number of features as a function of their reduced number of features. As can be seen, when the classifier learned the *raw* Fisher Vector features it achieved a UAR score of 63.98% in the CV. On the test set the performance was higher (66.12%). PCA proved to be useful here by contributing to a better classification performance in both CV and test phases (64.72% and 67.65%, respectively). However, we found that applying PN before PCA was effective as the CV and test UAR scores increased to 64.92% and 67.81%, respectively. Afterwards, we used the ComParE BoAW [22] feature set posterior probabilities and we combined them with those of the (power-normalized and reduced) Fisher Vectors, that is, we carried out a *late fusion*. The UAR score rose to 70.17% of UAR score on test set, which outperformed the BoAW baseline.

6 Conclusions

In this study, we presented the Fisher Vector approach as a method of classifying speech from subjects having a cold. Compared with studies done by other teams using the same dataset [11, 22], our performance is competitive. Moreover, our feature extraction approach seems to be simpler than that of the mentioned studies as we utilized one single type of feature representation for training a model. We found

that SVM gave better results when the feature pre-processing step was applied before executing the training phase. Thus, we demonstrated how applying Power Normalization along with dimension reduction via Principal Component Analysis on the Fisher Vector features improved the classification performance. Combining Power Normalization with PCA gave a better UAR score on test set. These results are higher compared to those got using the Bag-of-Audio-Words approach described in [22]. We can therefore say that PCA with the SVM allowed us to carry out a better classification of the actual data while taking care of the memory consumption. PN helped to reduce the impact of the features that increase their sparsity as the number of Gaussian components increase. Furthermore, L2-normalization was applied before fitting the data. This helped to alleviate the effect of having different utterances with distinct amounts of background information projected into the extracted features, which attempts to improve the prediction performance. In a future study, we will try out the FV approach on bigger datasets and evaluate the performance of a time-delay neural network when it uses them as input features.

References

[1] Arandjelovic, Relja and Zisserman, Andrew. All about VLAD. In *Proceedings of CVPR*, pages 1578–1585, 2013. DOI: 10.1109/CVPR.2013.207.

[2] Chang, Chih-Chung and Lin, Chih-Jeh. LIBSVM: A library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:1–27, 2011. DOI: 10.1145/1961189.1961199.

[3] Chatfield, Ken, Lempitsky, Victor, Vedaldi, Andrea, and Zisserman, Andrew. The devil is in the details: An evaluation of recent feature encoding methods. In *British Machine Vision Conference*, volume 2, pages 76.1–76.12, 11 2011.

[4] Cummins, N., Epps, J., Sethu, V., and Krajewski, J. Variability compensation in small data: Oversampled extraction of i-vectors for the classification of depressed speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 970–974, 2014. DOI: 10.1109/ICASSP.2014.6853741.

[5] Egas-López, José Vicente, Orozco-Arroyave, Juan Rafael, and Gosztolya, Gábor. Assessing Parkinson's Disease From Speech Using Fisher Vectors. *Proceedings of Interspeech*, pages 3063–3067, 2019. DOI: 10.21437/Interspeech.2019-2217.

[6] Egas-López, José Vicente, Tóth, László, Hoffmann, Ildikó, Kálmán, János, Pákáski, Magdolna, and Gosztolya, Gábor. Assessing Alzheimer's Disease from Speech Using the i-vector Approach. In *Proceedings of SPECOM*, pages 289–298. Springer, 2019. DOI: 10.1007/978-3-030-26061-3_30.

[7] Gosztolya, Gábor, Bagi, Anita, Szalóki, Szilvia, Szendi, István, and Hoffmann, Ildikó. Identifying schizophrenia based on temporal parameters in spontaneous

speech. In *Proceedings of Interspeech*, pages 3408–3412, Hyderabad, India, Sep 2018. DOI: 10.21437/Interspeech.2018-1079.

[8] Gosztolya, Gábor, Busa-Fekete, Róbert, Grósz, Tamás, and Tóth, László. DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification. In *Proceedings of Interspeech*, pages 3522–3526, Stockholm, Sweden, Aug 2017. DOI: 10.21437/Interspeech.2017-905.

[9] Gosztolya, Gábor, Grósz, Tamás, Szaszák, György, and Tóth, László. Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis. In *Proceedings of Interspeech*, pages 2026–2030, San Francisco, CA, USA, Sep 2016. DOI: 10.21437/Interspeech.2016-956.

[10] Gosztolya, Gábor, Grósz, Tamás, and Tóth, László. General utterance-level feature extraction for classifying crying sounds, atypical & self-assessed affect and heart beats. In *Proceedings of Interspeech*, pages 531–535, Hyderabad, India, Sep 2018. DOI: 10.21437/Interspeech.2018-1076.

[11] Huckvale, Mark and Beke, András. It sounds like you have a cold! Testing voice features for the Interspeech 2017 Computational Paralinguistics Cold Challenge. In *Proceedings of Interspeech*, pages 3447–3451. International Speech Communication Association (ISCA), 2017. DOI: 10.21437/Interspeech.2017-1261.

[12] Jaakkola, Tommi S. and Haussler, David. Exploiting generative models in discriminative classifiers. In *Proceedings of NIPS*, pages 487–493, Denver, CO, USA, 1998.

[13] Jolliffe, Ian. T. *Principal Component Analysis*. Springer-Verlag, 1986. DOI: 10.1007/978-1-4757-1904-8.

[14] Kaya, Heysem and Karpov, Alexey A. Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold. In *INTERSPEECH*, pages 3527–3531, 2017. DOI: 10.21437/Interspeech.2017-653.

[15] Lemaître, Guillaume, Nogueira, Fernando, and Aridas, Christos K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*, 18(1):559–563, 2017.

[16] Perronnin, F. and Dance, C. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of CVPR*, 2007. DOI: 10.1109/CVPR.2007.383266.

[17] Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukáš, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlíček, Petr, Qian, Yanmin,

Schwarz, Petr, Silovský, Jan, Stemmer, Georg, and Vesel, Karel. The Kaldi speech recognition toolkit. *Proceedings of ASRU*, 01 2011.

[18] Rosenberg, Andrew. Classifying skewed data: Importance weighting to optimize average recall. In *Proceedings of Interspeech*, pages 2239–2242, 2012. DOI: 10.21437/Interspeech.2012-131.

[19] Sánchez, Jorge, Perronnin, Florent, Mensink, Thomas, and Verbeek, Jakob. Image classification with the Fisher Vector: Theory and practice. *International Journal of Computer Vision*, 105:222–245, 2013. DOI: 10.1007/s11263-013-0636-x.

[20] Schuller, Björn, Batliner, Anton, Steidl, Stefan, and Seppi, Dino. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011. DOI: 10.1016/j.specom.2011.01.011.

[21] Schuller, Björn, Steidl, Stefan, Batliner, Anton, Bergelson, Elika, Krajewski, Jarek, Janott, Christoph, Amatuni, Andrei, Casillas, Marisa, Seidl, Amdanda, Soderstrom, Melanie, et al. The Interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proceedings of Interspeech*, pages 3442–3446, 2017. DOI: 10.21437/Interspeech.2017-43.

[22] Schuller, Björn, Steidl, Stefan, Batliner, Anton, Hantke, Simone, Bergelson, Elika, Krajewski, Jarek, Janott, Christoph, Amatuni, Andrei, Casillas, Marisa, Seidl, Amanda, Soderstrom, Melanie, Warlaumont, Anne S., Hidalgo, Guillermo, Schnieder, Sebastian, Heiser, Clemens, Hohenhorst, Winfried, Herzog, Michael, Schmitt, Maximilian, Qian, Kun, Zhang, Yue, Trigeorgis, George, Tzirakis, Panagiotis, and Zafeiriou, Stefanos. The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, Cold & Snoring. In *Proceedings of Interspeech*, pages 3442–3446, Stockholm, Sweden, Aug 2017.

[23] Seeland, Marco, Rzanny, Michael, Alaqraa, Nedal, Wäldchen, Jana, and Mäder, Patrick. Plant species classification using flower images: A comparative study of local feature representations. *PLOS ONE*, 12(2):1–29, 02 2017. DOI: 10.1371/journal.pone.0170629.

[24] Smith, David C and Kornelson, Keri A. A comparison of Fisher vectors and Gaussian supervectors for document versus non-document image classification. In *Applications of Digital Image Processing XXXVI*, volume 8856, page 88560N. International Society for Optics and Photonics, 2013. DOI: 10.1117/12.2023329.

[25] Vedaldi, Andrea and Fulkerson, Brian. VLFeat: an open and portable library of Computer Vision algorithms. In *Proceedings of ACM Multimedia*, pages 1469–1472, 2010.