Jitter and Shimmer Measurements for Speaker Diarization

Abraham Woubie[1], *Jordi Luque*[2], *and Javier Hernando*[1]

[1] Universitat Politecnica de Catalunya, BarcelonaTech, Barcelona, Spain
[2]Telefonica Research, Edificio Telefonica-Diagonal, Barcelona, Spain
abraham.woubie.zewoudie@upc.edu,jls@tid.es,javier.hernando@upc.edu

Abstract. Jitter and shimmer voice quality features have been successfully used to characterize speaker voice traits and detect voice pathologies. Jitter and shimmer measure variations in the fundamental frequency and amplitude of speaker's voice, respectively. Due to their nature, they can be used to assess differences between speakers. In this paper, we investigate the usefulness of these voice quality features in the task of speaker diarization. The combination of voice quality features with the conventional spectral features, Mel-Frequency Cepstral Coefficients (MFCC), is addressed in the framework of Augmented Multiparty Interaction (AMI) corpus, a multi-party and spontaneous speech set of recordings. Both sets of features are independently modeled using mixture of Gaussians and fused together at the score likelihood level. The experiments carried out on the AMI corpus show that incorporating jitter and shimmer measurements to the baseline spectral features decreases the diarization error rate in most of the recordings.

Keywords: speaker diarization, spectral features, jitter, shimmer, fusion

1 Introduction

Speaker diarization is the process of segmenting and clustering a speech recording into homogeneous regions and answers the question "Who spoke when" without any prior knowledge about the speakers [1]. A typical diarization system performs three basic tasks: first, it discriminates speech segments from the non-speech ones; second, it detects speaker change points to segment the audio data and finally, it groups these segmented regions into speaker homogeneous clusters. Speaker diarization can be used in different applications such as speaker tracking and speech recognition [1].

The performance of a speaker diarization system largely depends on successful extraction of relevant speaker independent features. Although short-term spectral features are the most widely used ones for different speech applications, the authors in [2] show that long term features can be employed to reveal individual differences which can not be captured by short-term spectral features. The current state-of-the-art speaker diarization systems usually make use of short-term spectral features as representation of speaker traits[3]. However, the work

of [4] and [5] show that the performance of the state-of-the-art speaker diarization systems can be improved by combining spectral features with prosodic and other long-term features.

Jitter and shimmer measure fundamental frequency and amplitude variations, respectively. Previous studies have shown that these voice quality features have been successfully used in speaker recognition and emotion classification tasks. The work of [6] shows that adding jitter and shimmer voice quality features to both spectral and prosodic features improves the performance of a speaker verification system. The work of [7] also reports that fusion of voice quality features together with the spectral ones improves the classification accuracy of different speaking styles and conveys information that discriminates the different animal arousal levels. Furthermore, these voice quality features are more robust to acoustic degradation and noise channel effects [8].

Based on these studies, we propose the use of jitter and shimmer voice quality features for speaker diarization task as they can add complementary information to the baseline spectral features. The main contribution of this work is the extraction of jitter and shimmer voice quality features and their fusion with the spectral ones in the framework of speaker diarization task. The experiments are tested on AMI corpus [9], a multi-party and spontaneous speech set of recordings, and assessed in terms of speaker diarization error (DER).

This paper is organized as follows. An overview of voice quality features used in this work is presented in Section 2. Section 3 provides an overview of agglomerative hierarchical clustering of speakers followed by fusion of spectral and voice quality features in Section 4. Experimental results are presented in Section 5 and finally, conclusions of the experiments are given in section 6.

2 Voice-quality features

Although the dominant features for speaker diarization are MFCC, studies such as [4] and [5] show that long term features such as prosody can also be used in speaker diarization systems. Long term features are able to acquire phonetic, prosodic and lexical information which cannot be captured by spectral ones.

Jitter measures variations of the pitch in voice whereas shimmer describes variation of the loudness. Studies show that these voice quality features can be used to detect voice pathologies [10]. They are normally used to measure long sustained vowels where measured values above a certain threshold are considered as pathological voices. Studies show that voice quality features have been successfully used in speaker recognition and other speech technology researches. For example, the work of [10] reports that jitter and shimmer measurements provide significant differences between different speaking styles.

Although different estimations of jitter and shimmer measurements can be found in the literature, we focus only on the following three measurements called absolute jitter, absolute shimmer and shimmer apq3 encouraged by previous work of [6]. The work of [6] has shown that these three measurements provided better results in speaker recognition tasks than the other jitter and shimmer

estimations. The three voice quality measurements are extracted over 30ms frame length at 10ms rate by means of Praat software [11]. Then, we calculate the mean of each of the three measurements over a window length of 500ms at 10ms step to smooth out fundamental frequency estimation and synchronize with the short-term spectral features.

2.1 Jitter measurement

Jitter (absolute) is a cycle-to-cycle perturbation in the fundamental frequency of the voice , i.e. the average absolute difference between consecutive periods, expressed as:

$$\text{Jitter (absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{1}$$

where T_i are the extracted pitch period lengths and N is the number of extracted pitch periods.

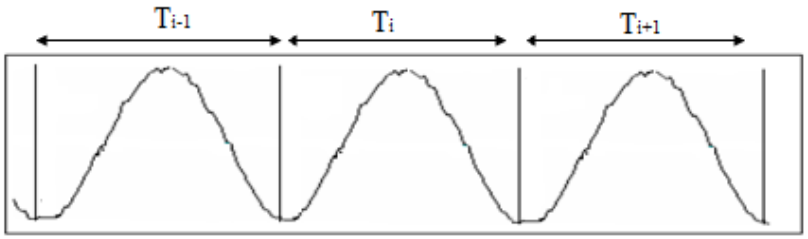

Fig. 1. *Jitter measurement for $N = 3$ pitch periods*

2.2 Shimmer measurement

– Shimmer (absolute) is the average absolute logarithm of the ratio between amplitudes of consecutive periods expressed as:

$$\text{Shimmer (absolute)} = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{2}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted pitch periods.

– Shimmer (apq3) is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbors, divided by the average amplitude expressed as:

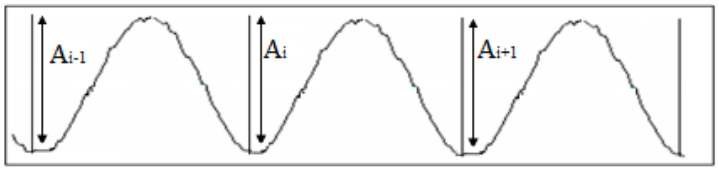

Fig. 2. *Absolute shimmer measurement for $N = 3$ pitch periods*

$$\text{Shimmer (apq3)} = \frac{1}{N-2}\sum_{i=2}^{N-1}\left|A_i - \left(\frac{A_{i-1}+A_i+A_{i+1}}{3}\right)\right| \tag{3}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted pitch periods.

3 Agglomerative hierarchical clustering of speakers

For this work, speaker diarization is performed on a mono-channel audio recording. Our approach is based on the bottom-up version of agglomerative hierarchical clustering(AHC). AHC-based diarization has been shown as one of the most successful approaches to address the problem of speaker clustering [12, 13]. Algorithm 1 highlights the main steps of the AHC popular technique. Input features $\{\mathbf{x}_i\}$ are partitioned in a set of segments $\hat{C}_i$, dividing the whole feature set. The clusters in the first iteration are initialized through previous segments and a model is built on them. Next, distances $d(\hat{C}_k, \hat{C}_l)$ among cluster models are computed in a pairwise comparison which aims to group similar regions. The initial clustering is iterated and the clusters are merged and aligned until some condition is fulfilled, e.g., a threshold on the previous distance matrix. Finally, each remaining cluster is expected to represent an ensemble of the data based on the selected distance measure.

Algorithm 1 Agglomerative Hierarchical Clustering (AHC), bottom-up alternative.

Require: $\{\mathbf{x}_i\},\ i = 1\ldots,\hat{n}:$ speech segments
$\hat{C}_i,\ i = 1,\ldots,\hat{n}$: initial clusters
Ensure: $C_i,\ i = 1,\ldots,n$: finally remaining clusters
1: $\hat{C}_i \leftarrow \{\mathbf{x}_i\},\ i = 1,\ldots,\hat{n}$
2: **repeat**
3: $i, j \leftarrow \text{argmin}\ d(\hat{C}_k, \hat{C}_l),\ k, l = 1,\ldots,\hat{n}, k \neq l$
4: merge $\hat{C}_i$ and $\hat{C}_j$
5: $\hat{n} \leftarrow \hat{n} - 1$
6: **until** no more cluster merging is needed
7: **return** $C_i,\ i = 1,\ldots,n$

In Figure 3, it is depicted a more detailed scheme of the AHC-based speaker clustering. The previous high level steps are adapted to the speaker diarization task jointly with the key idea that each cluster C_i should be composed exclusively by speech from the same speaker.

Speech activity detection(SAD): We have used Oracle SAD (the reference speech/non speech annotations) as our speech activity detection.

Cluster initialization: An initial segmentation is performed based on the homogeneous partition along time of the speech-only features, see (Fig. 3 block A). The number of initial clusters is selected automatically depending on the meeting duration but constrained to the range [35,65] clusters. It aims to deal with the trade-off between having a significant number of samples for modeling and avoiding common issues of AHC, such as overclustering and its high computational cost. So the number K_{init} of initial clusters is defined as

$$K_{\text{init}} = \frac{N}{G_{\text{init}}\, R_{\text{CC}}}, \tag{4}$$

where N stands for the number of features available per cluster and G_{init} is the number of Gaussians initially assigned to each cluster. The complexity ratio R_{CC}, the minimum number of frames per Gaussian, is fixed to 7 and the G_{init} to 5 Gaussians. Despite of its simplicity, this regular partition of the data allows the creation of "pure" enough initial cluster which is a key point in the algorithm [14, 15].

Acoustic modeling: Each set of acoustic features related to a cluster is independently modeled using HMM/GMM which is iteratively refined, (Fig. 3 block B). It is done in each clustering iteration through a two step training and decoding process. Each state of the HMM is composed by a mixture of Gaussians, fitting the probability distribution of the features by the classical expectation-maximization (EM) algorithm. Note that two independent HMM models are estimated per each feature stream but their log likelihoods given a feature are weighted as explained in Section 4. The number of mixtures is chosen as a function of the available seconds of speech per cluster in the MFCC features and fixed for the shimmer and jitter features. A time constraint, as in [16], is also imposed on the HMM topology which enforces the minimum duration of the speaker turn to be greater than 3 seconds.

Agglomerative distance is based on the Bayesian Information Criterion (BIC) as a metric among clusters. Furthermore, the stopping criterion, or ending point of the algorithm, is also driven by a threshold on the same matrix of distances, (Fig. 3 block C). A modified BIC-based metric [16] is employed to select the set of cluster-pairs candidates with smallest distances among them. Cluster-pair (i, j) is merged depending on whether its BIC_{ij} fulfills

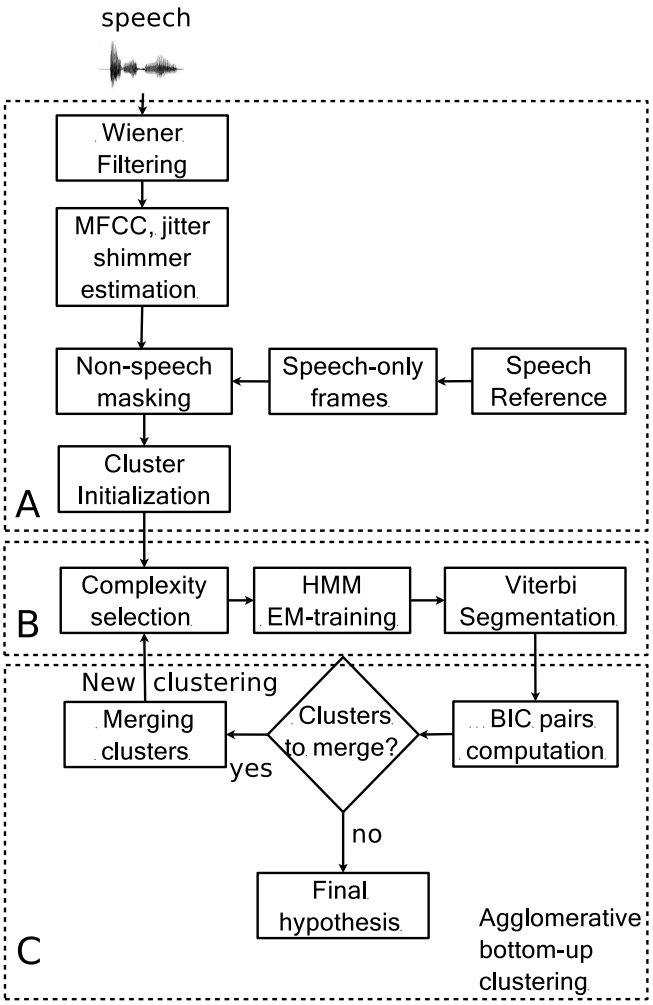

Fig. 3. *Speaker diarization scheme based on Agglomerative Hierarchical Clustering with automatic complexity selection.*

$$BIC_{ij} > \max\left(\gamma, BIC_{\mu} + \frac{3}{2} BIC_{\sigma}\right), \tag{5}$$

where BIC_{ij} is the BIC estimation between the clusters i and j performed as in [16] and γ is a threshold tuned on development data. The BIC_{μ} is the mean of BIC_{ij} for $i \neq j$ and the BIC_{σ} stands for the standard deviation of the same BIC set. Once clusters are merged, a two-step training and decoding iteration is performed again to refine the model statistics and align them with the speech recording, block B (see Fig. 3). The model complexity M_i^j, the number of

mixtures composing the model associated to cluster i at iteration j, is updated according to the R_{CC} value but only for the MFCC stream. In the case of voice quality features, Gaussian complexity is fixed manually and different values are explored. The automatic selection of the model complexity for MFCC features has shown a successful performance while it avoids the use of the penalty term in the classical BIC formulation [17, 12]. It is done by the following equation

$$M_i^j = \left\lfloor \left(\frac{N_i^j}{R_{\mathrm{CC}}}\right) + \frac{1}{2} \right\rfloor, \tag{6}$$

where N_i^j is the number of frames belonging to the cluster i. A more detailed description of the system can be found in [13, 18].

4 Fusion of spectral and voice quality features

Since the spectral and voice quality features have different dimensions and use different number of Gaussians per model, two independent HMM models have been estimated per each feature stream. The spectral features are used in parallel with voice quality features both in segmentation and clustering. The segmentation process uses the joint log likelihood ratio of both feature sets of the best path to create a segmentation hypothesis and the agglomerative clustering uses ΔBIC of fused Gaussian mixture mode scores to decide cluster merging. Given a set of input features vectors $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$, MFCC and jitter/shimmer respectively, which belong to same cluster, the log-likelihood is computed as a joint likelihood of both feature distributions as follows:

$$\log P(\mathbf{x}, \mathbf{y}|\theta_{ix}, \theta_{iy}) = \alpha \log P(\mathbf{x}|\theta_{ix}) + (1-\alpha) \log P(\mathbf{y}|\theta_{iy}), \tag{7}$$

where θ_{ix} is the model of cluster i using spectral feature vectors $\{\mathbf{x}\}$, and θ_{iy} is the model of the same cluster i using jitter and shimmer feature vectors $\{\mathbf{y}\}$. The weight of the spectral feature vector is α and, consequently, $(1-\alpha)$ is the weight of jitter and shimmer voice quality features.

5 Experiments

5.1 Database and experimental setup

The experiments are tested on AMI meeting corpus, a multi-party and spontaneous speech set of recordings, which consists of roughly 100 hours of speech. We have selected the 11 evaluation sets of the corpus to evaluate the diarization error rate of our approach. The average duration per meeting is around 27 minutes.

First of all, any noise of the input audio signal is minimized using Wiener filtering and we then apply speech activity detection algorithm to detect the speech segments and discard the non-speech ones.

Table 1. *Average DER results of the AMI corpus for different weighted combinations of MFCC, and Jitter and Shimmer features(JS) using 2 number of Gaussians for the JS.*

Feature set	**Weight of MFCC**	**Weight of JS**	**DER**
MFCC (Baseline)	1	0	24.76%
MFCC + JS	0.95	0.05	**21.45**%
MFCC + JS	0.925	0.075	22.76%
MFCC + JS	0.9	0.1	22.23%

The raw speech waveforms are then parameterized into sequences of MFCC using Fast Fourier Transform based log spectra with 30ms frame length and 10ms frame shift. The total number of coefficients extracted for the spectral features are 20. The extracted MFCC do not include deltas. The extraction of the three voice quality features is done as explained in Section 2. Fusion of the two set of features is done at the score likelihood level as explained in Section 4.

5.2 Experimental results

The performance of a speaker diarization system is evaluated using diarization error rate (DER) which represents the error contribution of missed speech, false alarm and speaker error. [1] We have used the reference speech/non speech annotations as our speech activity detection. The reason for using the reference speech/non speech annotations is that we are only interested to investigate the usefulness of voice quality features in reducing DER. The use of another speech activity detection may complicate the task and create more confusion. Therefore, the false alarms and missed speech have zero values in our experimental results.

As shown in Table 1, we have applied different weights for both features sets to find out the optimum set of weight values that provide us with the best results in terms of DER. The baseline system, which relies on spectral features, shows a DER of 24.76%. Weighting the MFCC by 0.95 and the voice quality features by 0.05 gives us a DER of 21.45%. It represents a 13.37% relative improvement compared to the baseline. We have observed that incorporatingg jitter and shimmer measurements to the baseline spectral features decreases the diarization error rate in nine of the eleven AMI recordings. Table 1 also shows that using different weight values for the jitter and shimmer features shows DER values better than the baseline.

We have also carried out an experiment to find out the best number of Gaussians for the voice quality features when its weight value is 0.05. The best DER result is found when we use 2 Gaussians as shown in Figure 4 which gives us a DER of 21.45% . The figure also shows that using one, three and five Gaussians provide better DER values than the baseline.The standard deviations of DER values in Figure 4 show the DER variations among the recordings.

[1] The scoring tool is the NIST RT scoring used as: ./md-eval-v21.pl -1 -nafc -o -R reference.rttm -S system_hypothesis.rttm

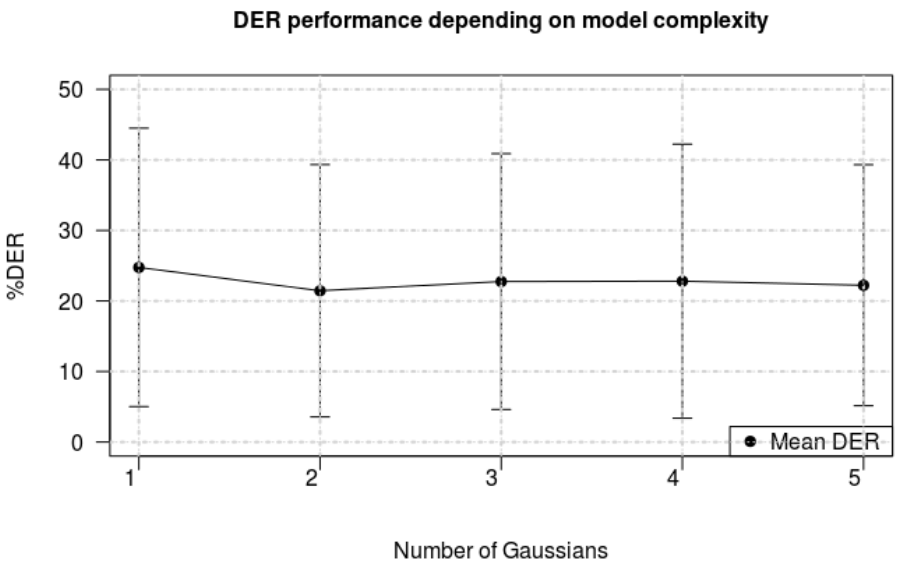

Fig. 4. *Diarization Error Rate (DER) and standard deviation as per the number of Gaussians for the JS with 0.95 and 0.05 weight values for MFCC and JS, respectively.*

6 Conclusions

We have proposed the use of jitter and shimmer voice quality features for speaker diarization experiment as these features add complementary information to the conventional baseline MFCC features. Jitter and shimmer voice quality features are first extracted from the fundamental frequency contour, and are then fused together with the baseline MFCC features. The fusion of the two streams is done at the score likelihood level by a weighted linear combination of the output log-likelihoods of each model. Our experiments show that fusing jitter and shimmer voice quality features with the baseline spectral features shows a 13.37% relative DER improvement.

7 Acknowledgements

This work has been partially funded by the Spanish Government projects TEC2010-21040-C02-01 and PCIN-2013-067.

References

1. Tranter, S. and Reynolds, D.: An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on* 14, 1557 –1565 (2006)

2. Farrús, M., Garde, A., Ejarque, P., Luque, J., and Hernando, J.: On the fusion of prosody, voice spectrum and face features for multimodal person verification. In *9th International Conference on Spoken Language Processing, ICSLP*, 2106–2109 (2006)
3. Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., and Vinyals, O.: Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech and Language Processing* (2011)
4. Friedland, G., Vinyals, O., Huang, Y., and Müller, C.: Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing* (2009)
5. Zelenák, M. and Hernando, J.: The Detection of Overlapping Speech with Prosodic Features for Speaker Diarization. In *INTERSPEECH*, 1041–1044 (2011)
6. Farrús, M., Hernando, J., and Ejarque, P.: Jitter and Shimmer Measurements for Speaker Recognition. In *INTERSPEECH* (2007)
7. Li, X., Tao, J., Johnson, M., Soltis, J., Savage, A., Leong, K., and Newman, J.: Stress and Emotion Classification using Jitter and Shimmer Features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, IV–1081–IV–1084 (2007)
8. Carey, M., Parris, E., Lloyd-Thomas, H., and Bennett, S.: Robust prosodic features for speaker identification. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3, 1800–1803 vol.3 (1996)
9. The Augumented Multi-party Interaction project,AMI meeting corpus. Website, `http://corpus.amiproject.org` (2011)
10. Kreiman, J. and Gerratt, B. R.: Perception of aperiodicity in pathological voice. *J Acoust Soc Am* 117 (2005)
11. Boersma, P. and Weenink, D.: Praat: doing phonetics by computer, Version 5.3.69. `http://www.praat.org/`
12. Fiscus, J. and et al.: The rich transcription evaluation project. Website, `http://www.nist.gov/speech/tests/rt/`
13. Luque, J. and Hernando, J.: Robust Speaker Identification for Meetings: UPC CLEAR-07 Meeting Room Evaluation System. In *Lecture Notes on Computer Science, LNCS*, vol. 4625. Springer-Verlag (2008)
14. Imseng, D. and Friedland, G.: Tuning-Robust Initialization Methods for Speaker Diarization. *Audio, Speech, and Language Processing, IEEE Transactions on* 18, 2028 –2037 (2010)
15. Luque, J., Segura, C., and Hernando, J.: Clustering initialization based on spatial information for speaker diarization of meetings. In *International Conference on Spoken Language Processing, ICSLP*, 383–386. Brisbane, Australia (2008)
16. Ajmera, J. and Wooters, C.: A robust speaker clustering algorithm . In *Proceedings of IEEE Speech Recognition and Understanding Workshop*. St. Thomas, U.S. Virgin Islands (2003)
17. Anguera, X., Wooters, C., and Hernando, J.: Robust Speaker Diarization for Meetings: ICSI RT06s evaluation system. In *International Conference on Spoken Language Processing, ICSLP* (2006)
18. Luque, J. and Hernando, J.: On the use of Agglomerative and Spectral Clustering in Speaker Diarization of Meetings. In *Odyssey 2012-The Speaker and Language Recognition Workshop* (2012)