On the Complementarity of Short-Time Fourier Analysis Windows of Different Lengths for Improved Language Recognition

Mireia Diez, Mikel Penagarikano, German Bordel,
Amparo Varona and Luis Javier Rodriguez-Fuentes

GTTS, Department of Electricity and Electronics,
University of the Basque Country, UPV/EHU, 48940, Leioa, Spain
mireia.diez@ehu.es

Abstract

Previous works have shown that remarkable performance improvements can be attained in speaker and language recognition tasks by combining several heterogeneous systems that provide complementary information. In this work, the complementarity of several i-vector language recognition systems, using Mel-Frequency Cepstral-Coefficient (MFCC) features computed on Short-Time Fourier Analysis windows of different sizes, is studied. Language recognition experiments carried out on the NIST 2007 and 2009 LRE datasets reveal relative performance gains of up to 33% when fusing the systems, with regard to the best single system. Results suggest that combining acoustic systems based on analysis windows of different sizes may allow to get advantage from both the sharper characterization of short events provided by short windows and the better frequency resolution of stationary events provided by long windows.

Index Terms: Spoken Language Recognition, Short-Time Fourier Analysis, i-vectors, Discriminative Score-Level Fusion.

1. Introduction

The combination of several heterogeneous systems, each providing complementary information, is known to provide remarkable performance improvements in verification tasks. The combination can be performed at early processing stages, concatenating the output features of different front-ends. It can be also performed at an intermediate stage, as in [1], where the i-vectors of two different prosodic systems were concatenated and plugged into a single classifier (in fact, this could be seen as an early stage combination, if the i-vectors were considered the output of a complex front-end). Finally, the combination can be performed at the score level by means of a discriminative fusion technique, such as Linear Logistic Regression [2, 3]. For example, in [4], thirteen heterogeneous subsystems developed by three different sites were combined into a single fused system. In [5], six subsystems based on different front-ends were combined to add robustness to severe noisy conditions.

Many speaker and language recognition systems rely on acoustic front-ends that extract the useful acoustic information and transform it into a compact representation. The well known Mel-Frequency Cepstral-Coefficients (MFCC) [6] and Perceptual Linear Predictive (PLP) [7] front-ends are typically applied to describe the quasi-stationary (slowly time varying) speech signals. Both front-ends estimate the instantaneous spectrum by means of short-time spectral analysis performed via the discrete Fourier Transform (DFT).

The main problem to be addressed when applying a DFT is leakage, which arises as a consequence of sampling and windowing. This problem is well studied and many window functions have been proposed to mitigate the undesired effects depending on the application requirements [8]. In any case, the use of windows that reduce leakage increases the variance. Some works addressed this problem too. Welch's modified periodogram [9] reduces the variance by time-averaging the spectrum estimates over multiple frames, but at the expense of loss in time resolution. The so-called Multi-taper technique, instead of forming a time-averaged spectrum, averages spectral estimates on a single frame, using a set of orthogonal window functions (a.k.a. *tapers*). Recent works have proven this multi-taper approach to improve a Speaker Recognition (SR) system [10, 11].

Another important aspect of the application of the DFT is the trade-off between time and frequency resolution. The choice of the window length (typically from 20 to 30 ms. for speech applications) is a compromise between the stationarity assumption and the frequency resolution defined as the ability to separate the contribution of two closed frequencies [12]. A wide window (i.e. a narrowband transform) ensures frequency resolution, but can violate the stationarity assumption. A narrower window (i.e. a wideband transform) supports the stationarity assumption but gives poor frequency resolution. In speech applications, the choice of the window length is a compromise between the expected set of speech sounds or phonemes comprising the target language. Vowels and voiced consonants could be expected to lie on wider sta-

tionary windows compared to other consonants. The selection of a single window length must look for a compromise between the requirements of the different units to cope with, and such a choice could lead to an information loss than could be avoided by using and combining different window sizes. In the case of Language Recognition (LR), where very different languages must be processed together, such an information loss could be more severe than in other applications.

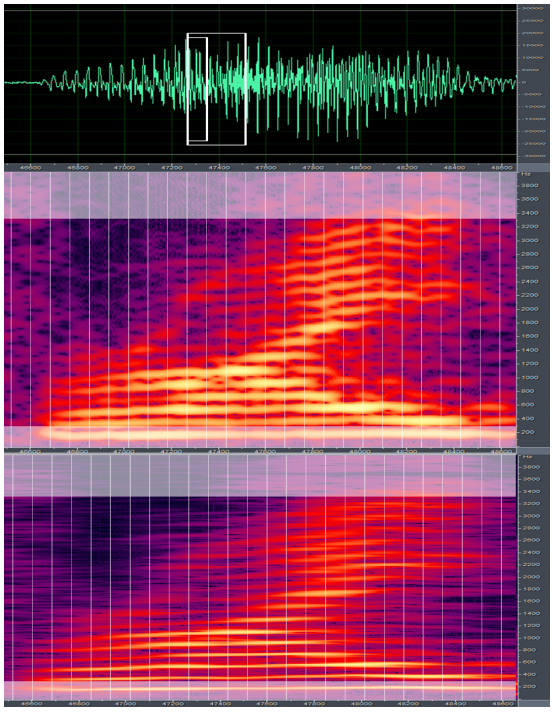

Figure 1: *Short-time Fourier analysis poses a trade-off on time and frequency resolutions. For a given signal (top), the use of 10ms. segments (middle) provides better time resolution but poorer frequency resolution than the use of 30 ms segments (bottom).*

In this paper, we study the complementarity of different window sizes for the short-time Fourier analysis of speech signals in a state-of-the-art acoustic LR system. Each front-end (with window sizes ranging from 10 to 30 ms) is plugged into a generative i-vector LR system [13] and the resulting systems are combined at the score level.

2. Feature Extraction

Frequency spectra were obtained from the 8KHz signals by applying 256-point DFT to Hamming windowed frames at a 10 ms frame rate. Mel filtering was thereafter applied between 300Hz and 3.3KHz to get 20 parameters.

The study was carried out considering five different window sizes. The standard window size used in Language Recognition is 20 ms. Therefore, values around it were selected for the experiments: 10, 15, 20, 25 and 30 ms. The minimum window size was set according to the standard window shift. Smaller window sizes would entail not analyzing part of the signal, thus losing information that could be useful.

An energy based VAD was applied, which removed frames with energies more than 30db below the maximum.

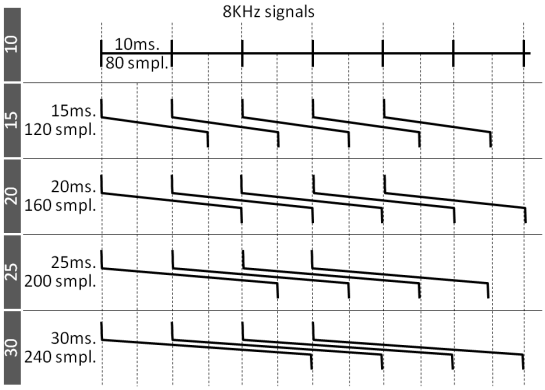

Figure 2: *Diagram of signal windowing using different lengths from 10ms. to 30ms.*

3. Experimental Setup

3.1. Datasets

3.1.1. NIST 2007 LRE

The NIST 2007 LRE [14] defined a spoken language recognition task for conversational speech across telephone channels, involving 14 target languages. For details on dataset distribution and configuration, see [15].

3.1.2. NIST 2009 LRE

The NIST 2009 LRE featured 23 target languages [16], involving 12 target languages for which Conversational Telephone Speech (CTS) was available in the NIST 2007 LRE dataset, plus 11 new target languages. For this competition, Broadcast Narrow-Band Speech was provided consisting mostly of telephone calls included in Voice of America (VOA) broadcasts. For details on the dataset partitions used in this work, see [15].

3.2. MFCC-SDC i-vector system

The concatenation of MFCC and Shifted Delta Cepstrum (SDC) coefficients under a 7-2-3-7 configuration was used as acoustic representation.

A gender independent 1024-mixture Gaussian Mixture Model (GMM) was used as Universal Background Model, (UBM), and estimated by Maximum Likelihood using the training set from each dataset. The total variability matrix (on which the i-vector approach relies) was estimated as in [17], using only target languages from the training sets. A generative modeling approach was ap-

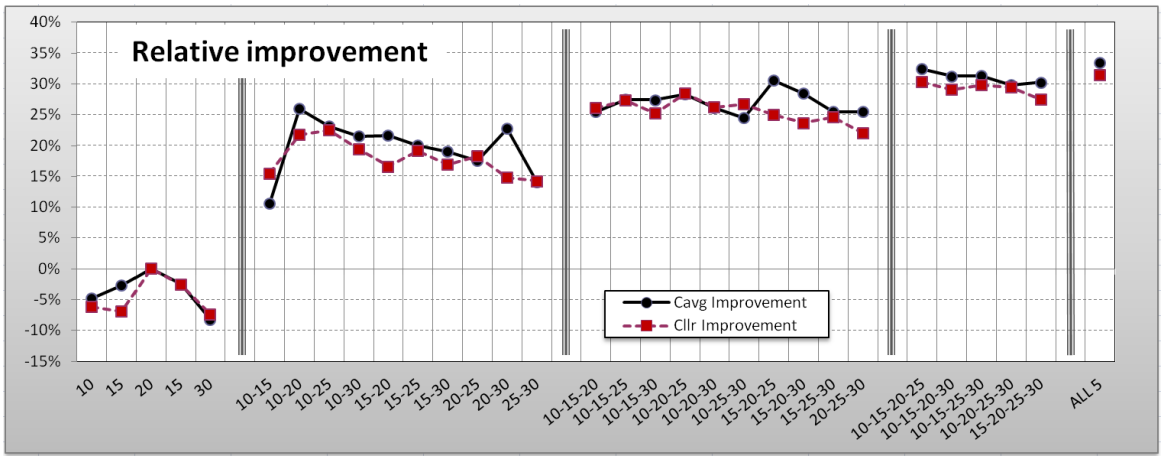

Figure 3: *Relative improvements attained by individual and fused systems on the NIST 2007 LRE primary task, taking the individual system that uses 20 ms window as reference.*

plied on the i-vector feature space (as in [13]), the set of i-vectors of each language being modeled by a single Gaussian distribution. Thus, the i-vector scores were computed as follows:

$$score(f, l) = N(w_f; \mu_l, \Sigma) \quad (1)$$

where w_f is the i-vector for target signal f, μ_l is the mean i-vector for language l and Σ is a common (shared by all languages) within-class covariance matrix.

3.3. Backend and fusion

A ZT-norm followed by a discriminative Gaussian backend was applied in experiments on both NIST 2007 and 2009 LRE datasets. The *FoCal* toolkit was used to estimate and apply the backend and calibration/fusion models [3].

3.4. Evaluation measures

In this work, systems are compared in terms of: (1) the average cost performance C_{avg} as defined in NIST evaluations up to 2009; and (2) the Log-Likelihood Ratio Cost C_{LLR} [3].

3.5. Statistical Significance

To measure the statistical significance of performance improvements (in terms of C_{avg}), a series of two-tailed paired t-tests was carried out [18], which gives an idea of the variability of performance improvements (and thus, the robustness of such improvements) across randomly defined sets of data. To that end, the NIST LRE evaluation datasets were split into 20 language-balanced disjoint random subsets. Then, C_{avg} values were computed on each subset for the best individual and the best fused systems (that is, the best pairwise, the best three-way fusion, etc.). Statistical significance tests were carried out using these performance figures.

4. Results and Discussion

Though no significant differences were found between individual system performances, the 20 ms system attained the best result among them. Smaller windows seemed to have slightly poorer performance.

Table 1: Actual $C_{\text{avg}} \times 100$ and C_{LLR} performance ranges for the MFCC-SDC-based i-vector systems and different fusions of them, on the NIST 2007 LRE primary evaluation task.

System: MFCC-SDC i-vector		$C_{\text{avg}} \times 100$	C_{LLR}
1 System		2.93 - 3.18	0.414 - 0.444
Fusion	2 Systems	2.17 - 2.62	0.321 - 0.355
	3 Systems	2.04 - 2.22	0.296 - 0.323
	4 Systems	1.98 - 2.06	0.288 - 0.300
	5 Systems	1.95	0.284

Table 1 shows performance ranges for individual systems and fusions of them on the NIST 2007 LRE primary evaluation task. For individual systems, performance ranges between 2.93 and 3.18 C_{avg}. When pairwise fusion is performed, results range from 2.63 C_{avg} to 2.17 C_{avg}. The statistical significance of the difference between the best individual system and the best pairwise fusion has been proven by performing a t-test. The fusion of three systems still provides a slight improvement, with the cost decreasing down to 2.04 C_{avg}. Fusions of more than three systems don't provide further significant improvements.

Figure 3 shows results for the individual systems and combinations of them on the NIST LRE 2007 primary task. The graphic reveals the significant relative improvements achieved by fusing systems. Overall, the fusion of the 5 individual systems attains a remarkable 33% relative improvement with regard to the best individual system, in terms of C_{avg}.

To further check the surprising complementarity found in the above reported experiments, a second series of experiments was carried out on the NIST 2009 LRE database. Table 2 shows performance ranges for individual systems and fusions of them on this dataset. Once again, fusions revealed a significant complementarity of systems using different window lengths. As for the 2007 LRE, pairwise fusions provided the most relevant improvement. In this case, the overall fusion of the systems reached a 18% relative improvement with regard to the best individual system in terms of C_{avg}.

Table 2: Actual $C_{\text{avg}} \times 100$ and C_{LLR} performance ranges for the MFCC-SDC-based i-vector systems and different fusions of them, on the NIST 2009 LRE primary evaluation task.

System: MFCC-SDC i-vector		$C_{\text{avg}} \times 100$	C_{LLR}
1 System		2.64- 2.69	0.536 - 0.550
Fusion	2 Systems	2.25 - 2.42	0.478 - 0.498
	3 Systems	2.13 - 2.30	0.462 - 0.481
	4 Systems	2.14 - 2.23	0.458 - 0.467
	5 Systems	2.17	0.457

4.1. Discussion

The above reported results prompted an inevitable question: Where is the complementarity coming from? Although our assumption was that using analysis windows of different lengths may provide complementary information by getting advantage from both, a sharper characterization of short events and a better frequency resolution of stationary events, it was necessary to discard other hypotheses.

In particular, the VAD applied in the experiments was dependent on the selected window length. Computing the frame energy on windows of different size, may imply filtering different frames in each system. Perhaps the fusion could be simply adding better information about where the speech was located, thus enhancing the overall performance.

To confirm or refute this hypothesis and to get a better insight on the origin of the attained gains, a new set of experiments was designed using an external VAD, so that VAD decisions were independent from the choice of window length. To that purpose, the open software Temporal Patterns Neural Network (TRAPs/NN) phone decoder for Hungarian, developed by the Brno University of Technology (BUT) [8], was used. VAD was performed by removing the feature vectors whose highest phone posterior value corresponded to the non-phonetic units (see [19] for details). Note that to properly use an external VAD, the analysis windows must be synchronized (centered) with the VAD flags. Therefore, in this approach some frames were discarded in each audio at the beginning and ending of each signal.

Table 3 shows performance ranges for individual systems and fusions of them on the NIST 2007 and 2009 LRE primary evaluation tasks, using an external VAD. Even though individual systems attain better performance than with the energy-based VAD, fusions of two systems still attain up to 25% and 8% improvements, with regard to the best individual system in terms of C_{avg} on the 2007 and 2009 datasets, respectively. The statistical significance of this improvement was checked with a t-test (see section 3.5 for details). Once again, fusions of 3 or more systems gave just some slight improvements. The overall performance gains attained, which reach 30% (2007 LRE) and 12% (2009 LRE), are still significant.

Table 3: Actual $C_{\text{avg}} \times 100$ and C_{LLR} performance ranges for the MFCC-SDC-based i-vector systems and different fusions of them, on the NIST 2007 and 2009 LRE primary evaluation task, using an external VAD.

NIST 2007 LRE			
System: MFCC-SDC i-vector		$C_{\text{avg}} \times 100$	C_{LLR}
1 System		2.55 - 2.85	0.374 - 0.407
Fusion	2 Systems	1.90 - 2.29	0.289 - 0.334
	3 Systems	1.78 - 2.05	0.272 - 0.304
	4 Systems	1.81 - 1.94	0.266 - 0.287
	5 Systems	1.78	0.262
NIST 2009 LRE			
System: MFCC-SDC i-vector		$C_{\text{avg}} \times 100$	C_{LLR}
1 System		2.49 - 2.80	0.515 - 0.558
Fusion	2 Systems	2.28 - 2.42	0.473 - 0.489
	3 Systems	2.23 - 2.31	0.463 - 0.474
	4 Systems	2.20 - 2.25	0.460 - 0.464
	5 Systems	2.19	0.457

5. Conclusions

In this paper, we have presented evidence of the complementarity of short-time Fourier analysis windows of different lengths. Experiments carried out have revealed that the fusion of different window-length based systems leads to significant improvements with regard to individual system performance in both NIST 2007 and 2009 LRE datasets.

The study suggests that performance gains may come from the complementarity of the sharper characterization of events provided by short-time windows and the better frequency resolution of stationary events attained with longer windows.

Future work will involve exploring the early fusion of features, or the intermediate (i-vector) level fusion of systems using analysis windows of different sizes.

6. Acknowledgements

This work has been supported by the University of the Basque Country under grants GIU13/28 and PES13/54.

7. References

[1] M. Kockmann, L. Ferrer, L. Burget, and J. Cernock, "iVector Fusion of Prosodic and Cepstral Features for Speaker Verification." in *Proceedings of Interspeech*. ISCA, 2011, pp. 265–268.

[2] N. Brummer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007.

[3] N. Brümmer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer, Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[4] L. J. Rodriguez Fuentes, M. Penagarikano, A. Varona, M. Diez, G. Bordel, A. Abad, D. Martinez, J. Villalba, A. Ortega, and E. Lleida, "The BLZ submission to the NIST 2011 LRE: Data Collection, System Development and Performance," in *Interspeech 2012*, Portland, Oregon, USA, 9-13 September 2012.

[5] A. Lawson, M. McLaren, Y. Lei, V. Mitra, N. Scheffer, L. Ferrer, and M. Graciarena, "Improving Language Identification Robustness to Highly Channel-Degraded Speech through Multiple System Fusion," in *Proceedings of Interspeech*. Lyon, Franec: ISCA, 2013.

[6] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, 1980.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 57, no. 4, pp. 1738–52, Apr. 1990.

[8] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[9] S. M. Kay, *Modern Spectral Estimation*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[10] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1990–2001, 2012.

[11] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Communication*, vol. 55, no. 2, pp. 237 – 251, 2013.

[12] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[13] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language Recognition in iVectors Space," in *Proceedings of Interspeech*, Firenze, Italy, 2011, pp. 861–864.

[14] A. F. Martin and A. N. Le, "NIST 2007 Language Recognition Evaluation," in *Proceedings of Odyssey 2008 - The Speaker and Language Recognition Workshop, paper 016*, 2008.

[15] M. Diez, A. Varona, M. Penagarikano, L. R. Fuentes, and G.Bordel, "On the Use of Phone Log-Likelihood Ratios as Features in Spoken Language Recognition," in *Proceedings of the IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, 2012.

[16] A. Martin and C. Greenberg, "The 2009 NIST Language Recognition Evaluation," in *Odyssey 2010 - The Speaker and Language Recognition Workshop, paper 030*, Brno, Czech Republic, 2010, pp. 165–171.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on ASLP*, vol. 19, no. 4, pp. 788–798, May 2011.

[18] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Porc. ICASSP*, 1989.

[19] M. Diez, A. Varona, M. Penagarikano, L. R. Fuentes, and G.Bordel, "Dimensionality Reduction of Phone Log-Likelihood Ratio Features for Spoken Language Recognition," in *Proceedings of the Interspeech 2011*, Lyon, France, 2013.