TIME-FREQUENCY REASSIGNED CEPSTRAL COEFFICIENTS FOR PHONE-LEVEL SPEECH SEGMENTATION

Georgina Tryfou[*†]*, Marco Pellin*[*] *and Maurizio Omologo*[*]

[*] Fondazione Bruno Kessler-irst
Via Sommarive, 18
38123 Povo, Italy

[†] ICT Doctoral School, University of Trento
Via Sommarive, 9
38123 Povo, Italy

ABSTRACT

This paper studies feature extraction within the context of automatic speech segmentation at phonetic level. Current state-of-the-art solutions widely use cepstral features as a front-end for HMM based frameworks. Although the automatic segmentation results have reached the inter-annotator agreement, within a tolerance equal or higher than 20ms, the same is not true when a lower tolerance is considered. We propose a new set of cepstral features that derive from the time-frequency reassigned spectrogram and offer a sharper representation of the speech signal in the cepstral domain. The features are evaluated through a series of forced alignment experiments which demonstrate a better performance, compared to the traditional MFCC features, in aligning phone boundaries within a small distance from their true position.

Index Terms— feature extraction, reassigned spectrogram, phonetic segmentation, forced alignment, HMM

1. INTRODUCTION

The accurate segmentation and labelling of speech into phone units is useful for diverse purposes, as for example the initialization of speech recognizers, the creation of databases for concatenative text-to-speech systems, the health related assessment of speech, and the evaluation of the performance of speech recognition tasks. The most accurate method of creating time-aligned phonetic labels is to employ an expert human annotator. This approach however, is expensive and requires an excessive amount of time. Moreover, the variability in human annotations results into subjective and unreproducible segmentation choices. Therefore, the design and implementation of automatic methods for phone-level segmentation of speech is of great interest.

Many different approaches have been exploited for addressing the task of automatic alignment, with most being based on either Hidden Markov Model (HMM), or Dynamic Time Warping (DTW). The latter primarily uses fixed templates, while in general HMM based approaches are characterised by more flexibility and provide superior results [1]. Therefore, HMM is the dominant technique in automatic segmentation of speech. In such systems, the acoustic signal and the phone transcriptions are used as input to a phone HMM based forced alignment system. In other words, the Viterbi algorithm is used for a constrained search of the phoneme boundaries inside the utterance, given the corresponding phonetic transcription. The segmentation results are evaluated as the percentage of correctly aligned boundaries, within different thresholds of tolerance. Because in continuous speech boundary positioning is an inherently subjective task, the goal of automatic phone alignment is often described as achieving the agreement between different human annotators. Within a tolerance of 20ms, the automatic methods have reached the 93.49% of inter-annotator agreement that has been reported in [1] for TIMIT dataset [2]. Nevertheless, when a lower tolerance is considered, the performance of automatic methods is still far from the corresponding inter-annotator agreement, that has been reported as high as 63% within 5ms for a dataset of German sentences [3].

A reason that contributes to the loss of accuracy with lower tolerances is related to the features used in the forced alignment systems. Cepstral features, for example Mel-frequency cepstral coefficients (MFCC) [4] and Perceptual Linear Predictive coefficients (PLP) [5], are currently the most popular choice [6, 7]. Both sets of features are obtained from the power spectrum as computed by the windowed speech signal. However, the application of the Short Time Fourier Transform (STFT) can be considered as a source of uncertainty as it suffers from a smearing effect and causes an unavoidable trade-off between temporal and spectral resolution. We therefore propose the use of the reassigned spectrogram [8, 9] in order to obtain a set of acoustic features which improve the accuracy of boundary positioning in forced alignment systems. The reassigned spectrogram provides an estimation of the instantaneous frequency of the input signal and, therefore, a more accurate representation of the time-frequency distribution of the energy.

The remainder of this paper is organized as follows. In section 2, we discuss in detail the most common acoustic features used in forced alignment systems. The method of time-frequency reassignment, and its incorporation in the calculation of a new set of acoustic features, are introduced in

section 3. In section 4, we describe the experimental activities, while the conclusions and future steps are presented in section 5.

2. MFCC AND PLP FEATURES

As mentioned above, cepstrum based features, namely MFCC and PLP, are the most popular choices as a front-end for speech segmentation systems, that employ an HMM based architecture. PLP cepstral features have been reported to be more robust in cases where a mismatch between the training and the testing material exists, while MFCC features have been found to perform better under clean and match conditions. There have been attempts to combine the most interesting characteristics of the two sets of features [10, 11], showing that the computation method of both can be further improved.

The block diagrams of the extraction steps for the two sets of acoustic features are presented in Figure 1(a). As depicted there, the processing is highly comparable. Both sets derive from the application of the STFT on the acoustic signal and the computation of the magnitude of each frequency bin. This results in the complete loss of the phase information, as well as in a possible loss of accuracy in the power spectrum estimation. In the case of MFCC, a pre-emphasis filtering is applied on the time-domain. In the case of PLP, the pre-emphasis takes place in the spectrum domain, according to an equal-loudness function. The subsequent frequency band analysis comprises the application of a Mel filter-bank in the MFCC computation and a Bark filter-bank in the PLP computation. Both scales, Mel and Bark, are perceptually inspired and in practice the differences between the resulting filter-banks are negligible. Higher frequencies components are emphasised and more filters are allocated for the lower frequencies. Concerning the intensity law (PLP) and the logarithmic compression (MFCC), both stages model the non-linear relation between the intensity of the sound and its perceived quality. The result of the two approaches has again a very similar effect.

It is therefore reasonable to state that a source of discrepancy between the two analyses is the method selected to map the spectrum into the cepstrum domain. MFCC are computed with the application of an inverse discrete cosine transform (IDCT) on the log Mel filter-bank output, a step that aims to the decorrelation of the features. In PLP analysis the auditory warped filter-bank output is further processed with inverse Discrete Fourier Transform, a step that yields the autocorrelation function. The values of the autocorrelation function are needed to compute the parameters of an LP model, which approximates the spectrum of the signal. Finally, cepstral coefficients are obtained from the LP model parameters, in a recursive fashion.

In both PLP and MFCC analysis, it is common practice to extend the feature vectors with their first and second order derivatives, in order encode their dynamic properties. This is usually performed with the application of a simple regression formula, that considers a certain number of neighbouring values. Finally, the features are very often normalized, for example employing cepstral mean normalization (CMN), variance normalization or band-pass filtering.

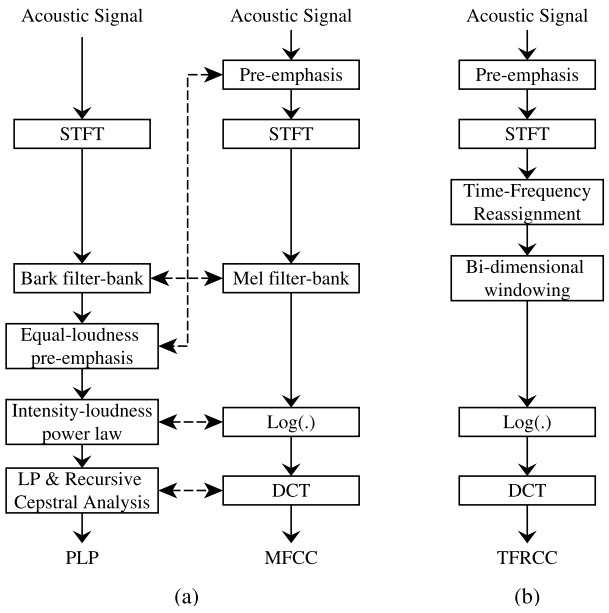

Fig. 1: The steps of extracting (a) PLP (left) and MFCC (right) features and (b) the proposed features. The dashed arrows indicate the analogous processing steps.

3. TIME-FREQUENCY REASSIGNED CEPSTRAL COEFFICIENTS

The time-frequency reassignment method, introduced in [8], remaps the spectral energy of each time-frequency point to a point closer to the actual region of support of a signal analysed with the STFT. When applied to speech signals the reassigned spectrogram offers a better localization of spectral features, such as the formant positions and the structure of the harmonics [12]. In spite of these interesting properties, the method of reassignment has been rarely used for speech processing [13]. Clear advantages were observed when applied in other fields, as for instance in music processing [14], though with a different perspective and implementation than the one proposed in this work. Here, we exploit the use of the reassigned spectrogram in the calculation of a set of time-frequency reassigned cepstral coefficients (TFRCC) which are adequate as a front-end to an HMM based speech segmentation system.

3.1. The time-frequency reassignment

The mathematical formulation of time-frequency reassignment is as follows. We denote $X(t, \omega)$ the continuous time

STFT of a signal, presented in the polar form as in

$$X(t,\omega) = M(t,\omega)e^{j\phi(t,\omega)} \quad , \tag{1}$$

where $M(t,\omega)$ is the magnitude and $\phi(t,\omega)$ is the phase of $X(t,\omega)$, defined as a function of continuous time t and angular frequency ω. The method of reassignment assigns to (t,ω) a new time-frequency coordinate that better reflects the distribution of energy in the analysed signal. The reassigned time-frequency coordinates $(\hat{t},\hat{\omega})$ may be calculated from the derivatives of the spectral phase as follows

$$\hat{t}(t,\omega) = -\frac{\vartheta\phi(t,\omega)}{\vartheta\omega} \tag{2}$$

$$\hat{\omega}(t,\omega) = \omega + \frac{\vartheta\phi(t,\omega)}{\vartheta t} \quad . \tag{3}$$

3.2. The feature extraction

The method for the extraction of the TFRCC features, as depicted in Figure 1(b), is performed in the following steps:

1. A pre-emphasis filter is applied to the speech signal.
2. The discrete STFT is calculated in order to obtain a complex spectrum. In the following, $X_h(t,\omega)$ denotes the discrete STFT of a signal, calculated with the use of an analysis window $h(n)$, that is shifted in time with a certain step.
3. In the case of the discrete STFT, the reassignment operations in (2) and (3) cannot be directly computed. Nevertheless, in [15] it is shown that the reassignment operations can be performed with the use of two auxiliary windows, as follows

$$\hat{t} = t - \Re\left\{\frac{X_{\mathcal{T}h}(t,\omega)\cdot X_h^*(t,\omega)}{|X_h(t,\omega)|^2}\right\} \tag{4}$$

$$\hat{\omega} = \omega + \Im\left\{\frac{X_{\mathcal{D}h}(t,\omega)\cdot X_h^*(t,\omega)}{|X_h(t,\omega)|^2}\right\} \quad , \tag{5}$$

   where $X_{\mathcal{T}h}$ is the discrete STFT computed using an analysis window, which is a time weighted version of $h(n)$, and $X_{\mathcal{D}h}$ is the discrete STFT computed using an analysis window, which is a frequency weighted version of $h(n)$. In practice, (4) and (5) reallocate spectral energy from the coordinate (t,ω) to the coordinate $(\hat{t},\hat{\omega})$ which can be formulated as

$$X(\hat{t},\hat{\omega}) = |X_h(t,\omega)|^2 \quad , \tag{6}$$

   with $X(\hat{t},\hat{\omega})$ defined in the continuous time-frequency domain. As a result, the estimates of the spectral energy distribution of the input speech signal are more precise.
4. The representation in (6), defined in the continuous time-frequency domain, cannot be directly used in the subsequent processing. In order to obtain a discrete version of $X(\hat{t},\hat{\omega})$ in a new time-frequency domain, a bi-dimensional window is applied. Since $X(\hat{t},\hat{\omega})$ is defined only at the points where there is energy to reassign, this new representation can be expressed as

$$S_w(m,k) = \sum_{(\hat{t},\hat{\omega})} w_k(m-\hat{t},\hat{\omega})X(\hat{t},\hat{\omega}) \quad , \tag{7}$$

   where $S_w(m,k)$ strongly depends on $w_k(\hat{t},\hat{\omega})$, which is a bi-dimensional window defined in the continuous time-frequency domain, m denotes the generic time instant in the new discrete time domain, and k denotes the index of a frequency range. Different weighting schemes can be exploited for the design of the window, which becomes more evident when it is expressed as

$$w_k(\hat{t},\hat{\omega}) = l(\hat{t})g_k(\hat{\omega}) \quad . \tag{8}$$

   In the above notation, $l(\hat{t})$ can be viewed as a continuous time window, that is shifted with a certain step, and $g_k(\hat{\omega})$ as a set of bandpass filters, for example a Mel-scale filter-bank, as the one used in MFCC, but defined in the continuous frequency space. The time resolution of the new time domain is determined by the length and the advance step of the time window $l(\hat{t})$, which should not be confused with the length and the advance step of the window $h(n)$ used for the calculation of the initial STFT. The frequency resolution is determined by the total number of filters in the filter-bank $g_k(\hat{\omega})$.
5. The discrete $S_w(m,k)$ is logarithmically compressed. The output of this step is essentially equivalent to the log mel-scale filter-bank output of MFCC, but it offers a better localization of the energy distribution of the signal.
6. The resulting representation is mapped into the cepstrum domain with the application of the IDCT, as typically done with MFCC.

Finally, common techniques, such as the augmentation of the vectors with time derivatives and the normalization of the cepstrum coefficients, can be applied on the TFRCC features.

4. EXPERIMENTAL RESULTS

For the evaluation of the proposed features we performed a set of speech segmentation experiments using forced alignment. The Hidden Markov Model Toolkit (HTK) was used to build phone HMMs, for which the probability estimates of the observations were modelled with Gaussian Mixture Models (GMM). The system was trained on the training partition of the TIMIT database (3696 read sentences, excluding the "sa" files) and tested in the full testing partition (1344 read sentences, excluding "sa" files). The complete set of 61 TIMIT phonemes was mapped into a set of 48 phonemes and each phonetic unit was represented by 4 states, as reported in [6]. The models were trained with the application of the Baum-Welch algorithm, with a total of 6 iterations over the data.

As a baseline configuration we used MFCC features, extracted with the following steps: (i) pre-emphasis of the

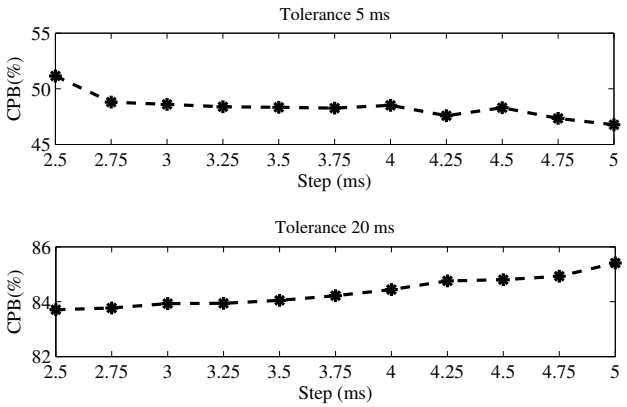

Fig. 2: Percentage of correctly positioned boundaries (CPB) for increasing advance step of a triangular time window of length 20ms, given a tolerance of 5 and 20ms.

frames with a pre-emphasis coefficient $\alpha = 0.97$, (ii) application of a 20ms Hamming window, , (iii) computation of the power spectrum with an analysis step size of 5ms, (iv) frequency warping with a Mel-scale filter-bank comprising 32 filters, (v) conversion to the logarithmic domain, (vi) application of the IDCT transform to obtain 12 cepstra coefficients and (vii) liftering of the cepstra to obtain a more narrow range of variances. CMN was applied and the log energy was added to the feature vector.

TFRCC feature vectors were extracted with the same configuration as above for (i) the calculation of the power spectrum of the acoustic signal, (ii) the pre-emphasis of the signal, and (iii) the application of the IDCT. CMN was applied and the log energy was added to the vector. For the design of the bi-dimensional window in (8), the same Mel-scale filter-bank as for MFCC was combined with an overlapping triangular window. It is noted here that alternative filter-bank configurations were explored and were found to have a similar effect as in the traditional cepstral features. The shape of the time window does not significantly affect the result. On the contrary, changes in the analysis step size result into more notable fluctuations, as presented in Figure 2.

Comparative segmentation results are reported in Table 1. For these experiments, the bi-dimensional window is created with a triangular time window of length 20ms, advancing in time with a step of 5ms. This, along with the 32-band filter-bank, produces the same time-frequency grid as in the case of the baseline MFCC configuration. The first two rows of Table 1 correspond to log-power spectrum domain feature sets, formed by the output of the Mel filter-bank in the case of MFCC features and the bi-dimensional windowing in the case of TFRCC features. The ability of the reassigned spectrogram to offer a more detailed representation of the fine structure of the time-frequency distribution of the acoustic signal is translated into a higher percentage of correctly aligned boundaries, particularly regarding low tolerances.

		Tolerance			
		5ms	10ms	15ms	20ms
Spectra	MFCC	36.22	64.63	78.42	84.43
	TFRCC	46.88	69.88	79.10	84.12
Cepstra	MFCC	37.55	65.21	79.12	85.09
	TFRCC	46.74	70.04	80.19	85.40
Δ, $\Delta\Delta$	MFCC	45.74	72.12	82.89	87.76
	TFRCC	49.82	73.26	82.86	87.40

Table 1: Percentages of correctly positioned boundaries, for different tolerances, using different feature sets.

The next two rows concern the results based on features derived from the application of the IDCT. MFCC-based segmentation presents improved results over all tolerance values. On the other hand, the TFRCC features demonstrate a different behaviour. In fact, a slight decrease of performance within 5ms indicates that the application of the IDCT is not the optimal choice for this step. Nevertheless, the boundary alignment improves when a tolerance higher than 10ms is regarded.

Finally, the last two rows of Table 1 are obtained by the extension of the feature set with the first and second order derivatives, considering a total of 3 and 7 frames, respectively. Focusing on the strictest threshold of tolerance, we observe that in the case of MFCC a relative improvement of 21.81% is presented. The corresponding improvement for TFRCC is 6.52%. This is explained by the fact that the TFRCC features are changing more rapidly than MFCC. Moreover, the use of the same regression formula, which is optimized for the MFCC features, fails to model the dynamic properties of the TFRCC. Nevertheless, the TFRCC features perform better, given a tolerance of 5 and 10ms.

It is also interesting to analyse the results with respect to transitions between different phonetic classes. In Table 2, we consider five phonetic classes: vowels, stops, nasals, fricatives and liquids. Both segmentation techniques demonstrate certain limitations in locating the boundaries in transitions such as vowel-to-vowel and liquid-to-vowel. This is expected since no unique point can be defined as boundary in such cases. In fact, such transitions in TIMIT database have been annotated with heuristic rules [2] which are not addressed in this experimental set-up. On the other hand, the TFRCC feature set presents an important improvement in better defined cases (36.5% relative improvement for any transition to vowel, 26.6% for any transition to fricative and 70.84% for any transition to liquid).

A final remark concerns the comparison of the results reported above to segmentation results reported in the literature, where results as high as 93.92% within a tolerance of 20ms have been reported in [7] for the TIMIT dataset. The experiments presented in this paper were designed to demonstrate the behaviour of the proposed features and compare them with

	vowel	stop	nasal	fric	liquid	all
vowel	**15.91**	52.52	**47.38**	40.56	14.02	39.32
	13.97	55.23	46.47	55.07	15.91	43.62
stop	42.82	42.18	29.54	37.06	29.21	39.73
	63.37	53.74	52.95	40.99	64.83	56.23
nasal	31.53	33.95	**20.00**	38.08	**28.41**	32.90
	51.57	34.02	17.50	44.31	27.84	42.49
fric	40.64	**50.12**	36.36	**32.76**	28.10	41.84
	55.37	49.14	53.11	29.80	53.60	52.16
liquid	**17.58**	45.66	**52.23**	36.20	19.08	23.33
	17.25	50.08	47.77	59.38	19.08	25.15
all	32.40	46.16	44.63	38.74	21.64	37.55
	44.23	51.66	47.00	49.07	36.97	46.74

Table 2: Percentage of correctly positioned boundaries per phonetic class within a tolerance of 5ms. For each transition pair, the first row corresponds to MFCC and the second to TFRCC. The transitions for which MFCC provide more accurate results (in bold) account for 24.7% of the testing material.

MFCC. All the results can be improved, as in [7], with the use of a more sophisticated HMM architecture, the use of context dependent models and the application of boundaries correction methods, which will be addressed in our future work.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a new set of features that can be used as a front-end for phone segmentation, as well as for speech recognition and other similar tasks. The proposed features result from the time-frequency reassigned spectrogram of the speech signal. In the experimental activities, they have been found to perform equally well compared with the traditional MFCC features, as far as more relaxed tolerance thresholds are concerned. On the other hand, they outperform MFCC features, with strict thresholds of tolerance. The power of the proposed feature set lies in the ability of the method of reassignment to offer a much sharper representation of the energy distribution of the speech signal. The experiments also indicated that further improvements are possible in the proposed analysis: in fact, both the application of the IDCT and the extension of the features with time derivatives do not yield an improvement as high as expected based on the behaviour of the forced alignment with MFCC features.

REFERENCES

[1] J. P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, vol. 51, no. 4, pp. 352–368, 2009.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium, 1993.

[3] M.-B. Wesenick and A. Kipp, "Estimating the quality of phonetic transcriptions and segmentations of speech signals," in *4th International Conference on Spoken Language*, 1996, pp. 129–132.

[4] S. 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, pp. 357–366, 1980.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, pp. 1738, 1990.

[6] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[7] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *14th Annual Conference of the International Speech Communication Association*, 2013, pp. 2306–2310.

[8] K. Kodera, R. Gendrin, and C. de Villedary, "Analysis of time-varying signals with small BT values," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 64–76, 1978.

[9] S. A. Fulop, Ed., *Speech Spectrum Analysis*, Springer, 2011.

[10] F. Hönig, G. Stemmer, C. Hacker, and F. Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *12th Annual Conference of the International Speech Communication Association*, 2005, pp. 2997–3000.

[11] B. Milner, "A comparison of front-end configurations for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, vol. 1, pp. I–797.

[12] F. Plante, G. Meyer, and W. Ainsworth, "Improvement of speech spectrogram accuracy by the method of reassignment," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 282–287, 1998.

[13] S. A. Fulop and Y. Kim, "Speaker identification made easy with pruned reassigned spectrograms," in *Proceedings of Meetings on Acoustics*. Acoustical Society of America, 2013, vol. 19.

[14] M. Khadkevich and M. Omologo, "Reassigned spectrum-based feature extraction for gmm-based automatic chord recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, pp. 1–12, 2013.

[15] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.