A Reassigned Front-End for Speech Recognition

Georgina Tryfou and Maurizio Omologo
Fondazione Bruno Kessler
Via Sommarive 18, Trento -Italy
Email: {tryfou,omologo}@fbk.eu

***Abstract*—This paper introduces the use of the TFRCC features, a time-frequency reassigned feature set, as a front-end for speech recognition. Compared to the power spectrogram, the time-frequency reassigned version is particularly helpful in describing simultaneously the temporal and spectral features of speech signals, as it offers an improved visualization of the various components. This powerful attribute is exploited from the cepstral reassigned features, which are incorporated in a state-of-the-art speech recognizer. Experimental activities investigate the proposed features in various scenarios, starting from recognition of close-talk signals and gradually increasing the complexity of the task. The results prove the superiority of these features compared to a MFCC baseline.**

I. Introduction

Automatic speech recognition (ASR) is based on statistical analysis of speech, performed by complex frameworks, for instance hidden Markov models (HMM) [1] or, more recently, deep neural networks (DNN) [2]. Before fed to such frameworks, the acoustic input signal is represented in compact form through sets of parameters, such as the Mel Frequency Cepstral Coefficients (MFCCs) [3] or the Perceptual Linear Predictive coefficients (PLPs) [4]. These are usually augmented with their first and second order derivatives [5]. The parametrization of the speech signals is designed to discard information that is considered irrelevant to the discrimination of the various speech units. Additional transformations aim at the reduction of the effects caused by environmental conditions, for example noise and reverberation, and the variabilities that exist among different speakers [6].

The common goal of the various approaches to the parametrization of speech, is to produce a compact set of values that describe the spectral shape of short segments of speech. Such segments are usually around 25ms long and are updated with a rate of around 10ms. Within each segment, the speech signal is assumed to be stationary, a fact that enables the use of the short-time Fourier analysis (STFT) for the estimation of the spectral content of the speech. Although the STFT enables the summarization of the speech content and the periodical update of the extracted parameters, there is a long list of alternative time-frequency distributions that have been studied in the context of speech processing [7].

Among these various time-frequency distributions, the time-frequency reassignment is a method that improves the representation of the speech spectral content, as it is very useful in representing simultaneously the temporal, *i.e.,*, onsets of plosive sounds, and the spectral features, *i.e.,*, harmonic structure of vowels, of speech signals [8]. In addition, when the recognized speech signal is impinged by reverberation, its spectral envelope, and therefore the MFCC features that describe this envelope, are smoothed and carry less information. The reassigned spectrogram, obtained from the method of time-frequency reassignment, is a sharpened version of the traditional spectrogram and the reassignment operation mitigates these smoothing disturbances that are introduced by the reverberation.

In [9], we proposed a set of cepstral features extracted from the time-frequency reassigned spectrogram of the speech signal, called Time-Frequency Reassigned Cepstral Coefficients (TFRCC), in order to address speech segmentation. TFRCC were proved particularly successful in detecting the boundaries between phones, when a very strict evaluation tolerance was considered. This can be attributed to the particularly good temporal resolution that can be achieved with the reassigned spectrogram, without sacrificing the spectral resolution. Here, we extend the scope of the work reported in [9], and investigate the TFRCC features when used as a front-end for speech recognition. We target various ASR scenarios, such as recognition of close-talk sentences, and of simulated and real reverberant versions of these sentences. In the distant speech recognition (DSR) task we further investigate how a front-end information fusion approach, namely channel selection (CS) can be combined with the proposed features.

The rest of this paper is organized as follows. In Section II we present an overview of front-end solutions for ASR. In Section III we overview the method of time-frequency reassignment and the TFRCC feature extraction method. In Section IV we present the experimental framework, setup and datasets, while the results of the conducted experiments are presented and discussed in Section V. Finally, we draw conclusions and discuss future activities in Section VI.

II. Related work

Feature extraction is the process of extracting sets of descriptors that represent specific properties of acoustic signals. Opposite to transformations, *e.g.* the Fourier transform, the feature extraction aims first, at representing higher level characteristics and, second, at significantly reducing the signal dimensionality.

Short-time frequency analysis has been extensively used in the majority of speech processing front-end techniques, since it was first introduced in 1940s [10]. Another important introduction in the field was the use of non-linear filter banks, as for instance those in the Bark and Mel scales [11], [12],

ISBN 978-0-9928626-7-1

as a means of modelling the nonlinear frequency resolution of human ear. The next advancement in feature extraction for ASR was the use of the cepstrum, which was introduced in [13] and then applied in the most commonly used feature sets until today, namely the MFCC and PLP.

An important part of front-end design for speech recognition are the various transformations that can be applied on certain feature vectors. Cepstral mean and variance normalization of cepstral features are both very common examples.Vocal tract normalization compensates for variabilities in the speech signal caused by the speaker dependent vocal tract shape.Similar effects can be achieved with the use of cepstral based linear transformations, such as Maximum Likelihood Linear Regression (MLLR) [14], and feature space MLLR (fMLLR) [15], commonly used for speaker adaptation.

Recent speech recognition evaluation campaigns, for example REVERB [16], CHiME-3 [17] and ASpIRE [18], indicate that state-of-the-art systems often choose sophisticated feature extraction methods, such as i-vector and gammatone ceptral coefficient [19], and incorporate additional front-end processing units, such as speech enhancement, beamforming and CS in order to improve recognition performance in real applications.

III. Time-frequency reassigned cepstral coefficients

The limitations of the short-time power spectrum that stem from the well known trade-off between time and frequency resolution have been extensively discussed in the literature [8], [20]. Time-frequency reassignment addresses this time-frequency trade-off, offering an improved representation of the temporal evolution of spectral components. Time-frequency reassignment has been exploited so far in the context of various applications, and utilized as the time-frequency representation of acoustic signals. Speech signal analysis and visualization is one of the most important application areas for the reassigned spectrogram. The suitability of the reassigned spectrogram in visualizing individual vocal chord pulsations has been very often exploited, as for example in [21]. In [22] the method of reassignment was applied in the context of speech formant analysis, and the notion of re-quantizing the reassigned spectrogram points at the STFT grid centers was introduced. In [23] the reassigned spectrogram was utilized in a double-vowel identification task which showed improvements over the recognition based on the traditional spectrogram. In a slightly different group of applications in the area of speech signal analysis, the reassigned spectrogram was exploited for speaker identification. The concept was first introduced in [24] and further discussed in [25].

A. Time-frequency reassignment

The polar form of the continuous time STFT of a signal is expressed as

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

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

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

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

The time-frequency reassigned point $(\hat{t},\hat{\omega})$ represents the center of gravity of the energy distribution of the signal. The method of reassignment results in a noisy representation, since random like noise appears in areas where there is no energy to reassign. Nevertheless, there is a de-noising technique that exploits a set of thresholds and can be used to address this problem [26].

B. Feature extraction

As proposed in [9], TFRCC features are extracted in a set of steps, similar to those used in the MFCC calculation. These steps are summarized as follows.

1) A pre-emphasis filter is applied to the speech signal.
2) The complex spectrum of the input, $X_h(t,\omega)$ is calculated, with the used of the discrete STFT.
3) In the case of the discrete STFT the reassignment operations in (2) and (3) cannot be directly computed, therefore the method described in [27] is used in order to reassign the spectrogram. In the obtained representation, $X(\hat{t},\hat{\omega})$, spectral energy from the coordinate (t,ω) has been reallocated to the coordinate $(\hat{t},\hat{\omega})$.
4) $X(\hat{t},\hat{\omega})$ is defined in the continuous time-frequency domain and has to be re-quantized in order to be exploited in the subsequent processing. This is a common step in applications that utilize the reassigned spectrogram as the time-frequency representation of data [28]. Instead of the common approach, that performs re-quantization by moving each reassigned point back to the closer STFT grid point, as in [29], we TFRCCs follow an approach that combines the re-quantization with the application of the Mel-scale filter-bank and an application of a moving window in time [9].
5) The discrete $S_w(m,k)$ is logarithmically compressed.
6) The features are mapped into the cepstrum domain with the application of the IDCT.

TFRCC features are essentially equivalent to the MFCC features, but they offer a better localization of the energy distribution of the signal.

IV. Recognition Experiments

The recognition experiments in this work were designed in order to investigate the behaviour of the TFRCC features compared to the MFCC features, under different acoustic conditions.

ISBN 978-0-9928626-7-1 © EURASIP 2017

A. Setup and datasets

For the experiments we used data from the DIRHA Project framework [30][1]. In particular, for the clean speech experiments we used close-talk recordings of the wall street journal (wsj) and phonetically rich (phrich) datasets, acquired in the FBK recording studio. Each of these sets comprises 409 sentences. For DSR experiments, we used data recorded in the living-room, a room with $T_{60} \approx 0.75s$, and we selected a set of 5 microphones installed on the walls and the ceiling of this room. For the training, we used the clean Wall Street Journal (WSJ0-5k) [31] training set. These utterances were reverberated with IRs measured in the living-room environment. As test material we used two different sets, extracted from the DIRHA-English corpus[2] [32], [33]. The first test set corresponds to the WSJ0-5k sub-set, and each of its two subsets *i. e.,* simulated and real, is composed of 409 sentences, uttered by 6 speakers. We call these sim-wsj and real-wsj respectively. The second test set is composed of phonetically-rich sentences, extracted from the Harvard Corpus, and it is called phrich dataset.

B. Recognition framework

1) Feature extraction: Each recognition experiment is performed for both sets of acoustic features, extracted from analysis frames of 25ms long, with an analysis rate of 10ms. Both sets of features are augmented with their first and second order derivatives. The TFRCC feature extraction is implemented within the Kaldi speech recognition toolkit [34], which is also used for building the recognizer.

2) Acoustic modelling: The "s5" Kaldi recipes concerning TIMIT and WSJ tasks were adapted to the DIRHA English framework as well as to model the front-end processing output described above. For our experiments, we consider 5 different acoustic models of increasing complexity. In the first level (mono), acoustic models represent 48 context independent phones. A three state left-to-right HMM is used to model each of the phones. The *tri1* acoustic models are based on simple triphone training, on features augmented with first and second order derivatives. After that, *tri2* and *tri3* acoustic models are trained on features transformed with linear discriminant analysis (LDA) and maximum likelihood linear regression (MLLR), with *tri3* models trained with speaker adaptive training. Furthermore, DNN running on top of the LDA-MLLR transformed features, were used. The DNNs were built according to Karel's recipe [35] with a network architecture shaped by 6 hidden layers of 1024 neurons, with a context window of 11 consecutive frames (5 before and 5 after the analysis frame), and an initial learning rate of 0.008.

3) Language modelling: Concerning the language modelling, for the wsj datasets we employ the bigram language model, as in the original Kaldi recipe. For the phrich dataset, in order to better focus on the behaviour of the proposed features in encoding acoustic information, we adopt a pure phone-loop as in [32]. Although this decision yields a loss in overall recognition performance, we avoid certain non-linear behaviours due to the language modelling.

[1]http://dirha.fbk.eu.

[2]The DIRHA-English dataset will be publicly distributed through the Linguistic Data Consortium (LDC)

TABLE I: Recognition WER results (%) for the clean WSJ dataset.

Features	mono	tri1	tri2	tri3	dnn
MFCC	22.9	11.1	10.4	6.3	3.7
TFRCC	22.7	11	10	5.8	3.5

TABLE II: Recognition PER results (%) for the clean phrich dataset.

Features	mono	tri1	tri2	tri3	dnn
MFCC	47.3	42.8	40.2	32.9	28.1
TFRCC	47.3	41.9	39.2	32.1	27.2

V. Experiments and results

A. Close-talk performance

Here, we report the recognition results that are obtained for the close-talk sentences of each dataset, as these were recorded in the FBK recording studio. The recognition results for the clean wsj test set are presented on Table I. Concerning the acoustic models, as expected the use of more complex models, from *mono* to *DNN* based ones, results in significant improvements on the recognition performance. In addition, we observe the consistent improvements that the TFRCC features yield, compared to the MFCC features, for all the studied acoustic model types.

Next, Table II reports the results for the close-talk recordings of the phrich utterances. The improvement of the recognition performance with the use of more complex acoustic models is still evident in this experiment. Finally, also for this dataset the TFRCC features result in improved recognition performances.

B. Performance under reverberation

Here, we study the performance of various features and acoustic models in reverberant conditions. First, we present the single distant microphone (SDM) results for the set of 5 microphones that was selected, as described in Section Section IV. The results reported in Table III correspond to the simwsj set and in Table IV to the real-wsj set. As expected, the presence of reverberation drastically reduces the recognition performance for both cases. Nevertheless, we still observe that the use of TFRCC features results in improvements of the performance evident in all the microphones considered here. It is interesting to note that for each feature set the microphone that corresponds to the lower word error rate (WER) is not always the same. For instance, in the last columns of Table 4, we observe that MFCC result in the lowest WER for microphone LIC, while TFRCC achieve the higher performance for microphone L4L. This trend may suggest that TFRCC and MFCC provide a different behaviour in the modelling of reverberated speech signals. However, we plan to conduct an in-depth analysis on this issue to better correlate

ISBN 978-0-9928626-7-1 © EURASIP 2017

TABLE III: SDM WER results (%) for the recognition of the sim-wsj dataset

(a) Results using MFCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	65.5	42.3	36	24.8	16.1
L2R	63.9	41.2	35.4	24.4	15.5
L3L	65.2	41.9	35.9	24.8	16.2
L4L	67.5	43.4	37	24.9	16.2
LA6	68.5	44.3	38.9	26.3	17.1
Avg	66.1	42.6	36.6	25.	16.2

(b) Results using TFRCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	63.9	41.3	34.7	23.9	15.6
L2R	64.2	39.5	35.1	24.2	15
L3L	63.3	40.2	34.4	23.5	15.7
L4L	65.2	41.4	35.4	24.2	15.9
LA6	66.1	42	37.1	25.4	16.5
Avg	64.5	40.9	35.3	24.2	15.7

TABLE IV: SDM WER results (%) for the recognition of the real-wsj dataset

(a) Results using MFCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	66.7	40.9	33.9	23.1	14.5
L2R	68.1	43.1	37	24.1	16.7
L3L	64.5	40.6	33.6	22.8	15.1
L4L	64.4	41.9	34.1	23.3	15.4
LA6	66.2	42.4	35.7	22.9	15.4
Avg	66	41.8	34.9	23.2	15.4

(b) Results using TFRCC based front-end

Mic	mono	tri1	tri2	tri3	dnn
L1C	65.1	40.5	34.2	22.8	14.9
L2R	67.5	42.9	35.8	24.1	16.5
L3L	64	38.9	33.2	22.3	14.4
L4L	64.2	41	33.4	22.7	14.2
LA6	65.4	40.6	33.6	22.6	14.4
Avg	65.2	40.8	34.1	22.9	14.9

this specific experimental evidence with the properties of the front-end processing.

Next, we study how the TFRCC features perform in a common setup for multi-microphone DSR, namely CS [36], [37]. According to this practice, from the set of available microphones only one is chosen to be used for the decoding of each utterance. Table V summarizes different CS results for the sim-wsj dataset, using DNN based acoustic models.

TABLE V: CS results (%) for the simdev-WSJ dataset. The reported results correspond to DNN based acoustic models.

	MFCC	TFRCC
Avg. SDM	16.2	15.7
Oracle	10.16	9.57
CD informed	13.69	13.5
CD blind	14.71	14.28

TABLE VI: Recognition PER results (%) for the reverberant phrich dataset.

Features	mono	tri1	tri2	tri3	dnn
MFCC	69.5	64	62.5	60.9	54.9
TFRCC	69.1	63.6	61.6	57	52.4

First, the average SDM results, as taken from the last rows of Tables III and IV can be considered a lower bound of a CS method. Next, an oracle selection of the best microphone is the upper bound of any CS method, since it performs an a posteriori selection of the best recognition output. Finally, we report the results of two actual CS methods, which use cepstral distances (CD) in order to perform CS. The first, CD informed, uses the close-talk reference while the second, CD blind, does not. More details on CD based CS can be found in [38]. We observe that the TFRCC features consistently result in improved recognition performance, not only for the upper and lower bounds of CS, but for both CD based approaches as well. These results reinforce the use of TFRCC features in the context of multi-microphone DSR.

VI. Conclusions

In this work we presented a set of experimental results for the recognition of clean and reverberant data, based on the use of a time-frequency reassigned set of features. We found that these features consistently lead to improvements, compared to the use of the MFCC features. Since the results are so far encouraging, we still wish to further investigate several aspects of the proposed TFRCC features. For instance, we are interested in understanding how the various thresholds that can be applied on the time-frequency reassigned spectrogram in order to remove some of the random noise, can affect the recognition results. These thresholds manage the amount of harmonic and impulsive information in the final representation, and therefore have an important role in the description of the spectral and temporal features of speech signals. In addition, we aim to the design of a CS method that is based on the particular characteristics of the reassigned spectrogram, and can further improve multi-microphone DSR.

References

[1] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] 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.

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

[5] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

ISBN 978-0-9928626-7-1 © EURASIP 2017

[6] J. Cohen, T. Kamm, and A. G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3246–3247, 1995.

[7] L. Cohen, *Time-frequency analysis*. Prentice hall, 1995, vol. 778.

[8] S. A. Fulop, *Speech Spectrum Analysis*. Springer Berlin Heidelberg, 2011.

[9] G. Tryfou, M. Pellin, and M. Omologo, "Time-frequency reassigned cepstral coefficients for phone-level speech segmentation," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 22nd European*. IEEE, 2014, pp. 2060–2064.

[10] W. Koenig, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," *The Journal of the Acoustical Society of America*, vol. 18, no. 1, pp. 19–49, 1946.

[11] E. Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[12] S. Stevens and E. Volkmann, J .and Newman, "The mel scale equates the magnitude of perceived differences in pitch at different frequencies," *J. Acoust. Soc. Am*, vol. 8, no. 3, pp. 185–190, 1937.

[13] A. M. Noll, "Short-time spectrum and "cepstrum" techniques for vocal-pitch detection," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 296–302, 1964.

[14] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[15] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[16] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, ser. WASPAA, 2013, pp. 1–4.

[17] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[18] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) Challenge," in *Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2015.

[19] M. J. Alam, V. Gupta, P. Kenny, and P. Dumouchel, "Use of multiple front-ends and i-vector-based speaker adaptation for robust speech recognition," *Proc. of REVERB Challenge*, pp. 1–8, 2014.

[20] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, no. 7, pp. 941–981, 1989.

[21] K. Fitz and S. A. Fulop, "A unified theory of time-frequency reassignment," *arXiv preprint arXiv:0903.3080*, 2009.

[22] 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.

[23] G. F. Meyer, F. Plante, and F. Berthommier, "Segregation of concurrent speech with the reassigned spectrum," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1203–1206.

[24] S. A. Fulop and S. F. Disner, "The reassigned spectrogram as a tool for voice identification," in *International Congress of Phonetic Sciences*, 2007, pp. 1853–1856.

[25] S. A. Fulop and Y. Kim, "Speaker identification made easy with pruned reassigned spectrograms," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. Acoustical Society of America, 2013.

[26] S. A. Fulop and K. Fitz, "Separation of components from impulses in reassigned spectrograms," *Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1510–1518, 2007.

[27] 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.

[28] S. Hainsworth, M. Macleod, S. W. Hainsworth, and M. D. Macleod, "Time frequency reassignment: A review and analysis," Cambridge University Engineering Department, Tech. Rep., 2003.

[29] F. Plante and W. A. Ainsworth, "Formant tracking using reassigned spectrum," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[30] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, and P. Maragos, "The DIRHA simulated corpus," *Proc. of International Conference on Language Resources and Evaluation*, vol. 5, may 2014.

[31] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "Continous speech recognition (CSR-I) Wall Street Journal (WSJ0) News Complete," *LDC93S6A. DVD. Linguistic Data Consortium, Philadelphia*, 1993.

[32] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-English corpus and related tasks for distant-speech recognition in domestic environments," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU, 2015, pp. 275–282.

[33] M. Ravanelli, P. Svaizer, and M. Omologo, "Realistic multi-microphone data simulation for distant speech recognition," in *Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH, 2016.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. o. Schwarz, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU, 2011.

[35] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[36] C. Guerrero, "Information fusion approaches for distant speech recognition in a multi-microphone setting," Ph.D. dissertation, University of Trento, 2016.

[37] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170–180, 2014.

[38] C. Guerrero, G. Tryfou, and M. Omologo, "Channel selection for distant speech recognition - exploiting cepstral distance," in *Annual Conference of the International Speech Communication Association*, ser. INTERSPEECH, 2016.

ISBN 978-0-9928626-7-1 © EURASIP 2017