An Efficient Front-End for Distributed Speech Recognition over Mobile

Addou Djamel, Selouani Sid Ahmed, Malika Boudraa, and Bachir Boudraa

Abstract—**To improve the robustness of distributed speech front-ends in mobile communication we introduce, in this paper, a new set of feature vector which is estimated through three steps. First, the Mel-Line Spectral Frequencies (MLSFs) coefficients are combined with conventional MFCCs, after extracted from a denoised acoustic frame using the wiener filter. Also, we optimize the stream weights of multi-stream HMMs by deploying a discriminative approach. Finally, these features are adequately transformed and reduced in a multi-stream scheme using Karhunen-Loeve Transform (KLT). Recognition experiments on the Aurora 2 connected digits database reveal that the proposed front-end leads to a significant improvement in speech recognition accuracy for highly noisy GSM.**

Index Terms—**Distributed speech recognition, front-end processing, mel-frequency coefficients, Mobile communications, noise robustness.**

I. Introduction

Environmental robustness is an important area of research in speech recognition. Mismatch between trained speech models and actual speech to be recognized is due to factors like background noise. It can cause severe degradation in the accuracy of recognizers which are based on commonly used features like Mel-Frequency Cepstral Coefficient (MFCC). It is well understood that all previous auditory based feature extraction methods perform extremely well in terms of accuracy due to the dominant frequency information present in them and they have become standard and currently used in systems for distributed speech recognition (DSR). For such systems, it is crucial to use robust features to maintain a good performance when the signal to noise ratio (SNR) decreases. In order to face this difficulty many techniques have been developed. They are centered upon two major approaches. The first approach aims at establishing a compensation method for clean models in order to adapt to new environments. The second approach aims at extracting, through a robust parameterization process, the relevant information while eliminating noises and artifacts. A broad range of techniques exists for conveniently representing the speech signal in mismatched conditions [8,11]. However, most of the current approaches assume that the speech and noise are additive in the linear power domain and the noise is stationary. In this paper we are concerned with the optimization of the parameterization process in order to maintain, in noisy conditions, the relevant part of the information within a speech signal while eliminating their noise-corrupted part for the DSR over GSM networks.

In previous work [1,13], we introduced a multi-stream paradigm for DSR in which, we merge different sources of information about the speech signal that could be lost when using only the MFCCs to recognize uttered speech. Our experiments showed that such multi-variable, integrating some parameters based on a model simulating the cochlea and the acoustic cues reflecting the spectral resonances (formants), leads to an improved recognition rate. This showed that the MFCC, despite their popularity, lose the appropriate information to the process of recognition in strongly noisy environment. We used a 3-stream feature vector. The first stream vector consists of the classical MFCCs and their first derivatives, whereas the second stream vector consists of acoustic cues derived from hearing phenomena studies. The magnitudes of the main resonances of the spectrum of the speech signal were used as the elements of the third stream vector. The above-mentioned work has been extended in [13] by the use of the formant frequencies instead of their magnitudes for ASR within the same multi-stream paradigm. In these experiments, the recognition of speech is performed using a 3-stream feature vector, which uses the formant frequencies of the speech signal obtained through an LPC analysis as the element of the third stream vector combined with the auditory-based acoustic distinctive features and the MFCCs. The obtained results showed that the use of the formant frequencies for ASR in a multi-stream paradigm improves the ASR performance. Then in [1], we extended our work to evaluate the robustness of the above mentioned proposed features using a multi-stream paradigm for ASR in distributed environments. In this latter configuration the weights of each stream is determined empirically and remain constant after being fixed. The obtained results showed that the use of such features renders the recognition process more robust in noisy environments of Aurora-2 tasks.

Many studies have been published in order to propose the robust recognition systems in mobile telecommunications [16]. In the proposed front-end the state-of-the-art MFCC features are supplemented by MLSFs features (Mel Lines Spectral Frequencies). It is important to note that MLSFs have the advantage of being used in systems for speech coding. The integration of MLSFs feature sets is done based on the multi-stream paradigm. Furthermore for optimizing the stream exponents (also known as weights) of multi-stream HMMs, a distinctive technique is proposed, by deploying a discriminative approach. On the other hand, we aim to optimize the use of flow parameters by reducing the

Manuscript received April 30, 2012; revised June 3, 2012.

D. Addou, M. Boudraa, and B. Boudraa are with the Faculty of Computer Sciences and Electronical at Sciences & Technology Houari Boumediene University, Algiers, Algeria (e-mail: daddou@ usthb.dz).

S. A. Selouani is with the School of Engineering and Mathematical Sciences at Moncton University, New Brunswick Canada (e-mail: selouani@umcs.ca).

size of acoustic vectors while improving system robustness. An effective way to perform this reduction is to use the Karhunen-Loéve transformation (KLT) [7]. This is a technique of decomposition into subspace also used in enhancement of noisy signals [6], [14]. Thus integrating KLT in our approach, we realize two objectives: reduction of optimal parameters and improved robustness, by eliminating the noisy principal components.

This work presents a complementary solution for the extraction of acoustic parameters adapted to a noisy environment. The rest of the paper is organized as follows. Section 2 gives an overview on distributed speech recognition and section 3 describes an alternative approach based on the combination of multiple feature sets was proposed. Section 4 is devoted to experimental validation and analysis of its results. A conclusion, on the present work, completes this article.

II. Overview of Distributed Speech Recognition

Transmitted speech over mobile channels can significantly degrade the performance of speech recognizers when compared to the unmodified signal. This is due to the low bit rate speech coding as well as channel transmission errors. One solution to these problems might be elimination of the speech channel and instead using an error protected data channel to send a parameterized representation of the speech, which is suitable for recognition. By doing this, the recognition process is distributed between the terminal and the network which is why such systems are known as distributed speech recognizers. The ETSI Aurora standard [3] was originally created for speech recognition on distributed architectures. The terminal has a charge of extracting cepstral parameters and transmits them after compression (Fig. 1). The compressed stream is then received by a remote server for recognition. Degradation due to coding of low debit voice or channel coding is avoided.

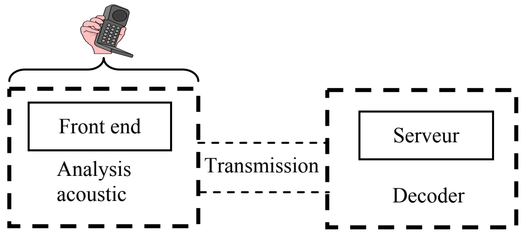

Fig. 1. Bloc diagram of DSR system.

The standard ETSI DSR-FE is mainly based on MFCC features. These features are the most popular features in current ASR systems. However, the performance of an MFCC cepstrum based system deteriorates in the presence of noise. In order to improve the noise robustness of the DSR front-end; one must combine the MFCCs with features that are robust against noise. PLP features [4], RASTA features [5], and spectral peaks, also known as formant-like features [1], are some of the features that are known to be robust against additive noise. Choosing the best feature set highly depends on the application and constraints. In DSR systems the feature extraction process takes place on a mobile set with limited processing power. On the other hand there is a certain amount of bandwidth available for each user for data transmission. Among the features mentioned above, MLSF features are more suitable for this application since; extracting them can be done as part of the process of extracting MFCCs, which saves a lot of computational process. Also, the two main reasons which have motivated our choice to consider the MLSFs in noisy communications mobile are; the first relates to the fact that the MLSF regions of the spectre can stay above the noise level even if the SNR is very low, while regions of lower energy tend to be masked by the energy of noise. The second reason is related to the fact that MLSFs are commonly used in conventional speech codec. This prevents the incorporation of new parameters that may require significant and costly changes to existing devices and codec.

III. Integration of Multiples Features Sets

The superscript numeral used to using HMMs with multiple streams has been adopted by many researchers [2]. This early-stage feature combination approach has the advantage of computational simplicity as well as implementation feasibility. In this method multiple acoustic feature streams obtained from different sources are concatenated to form a multi- stream feature set which is then used to train multi-stream HMMs.

Consider *S* information sources that provide time synchronous observation vectors O_{ts}; $s = 1...S$ at each time instant *t*. The dimensionality of the observation vectors can vary from one source to another. Each time sequence of the observation vector provides information about a sequence of hidden states j (j = *1... J*). In a multi-stream system, instead of generating *S* state sequences from *S* observation sequences, only one state sequence is generated. This is actually done by introducing a new output distribution function for states. The output distribution of state *j* is defined as:

$$b_j(o_t) = \prod_{s=1}^{S} [b_{js}(o_{st})]^{\gamma_{js}} . \qquad (1)$$

The exponent γ specifies the extent to which each stream contributes to the overall distribution by scaling the output distribution of each feature stream. The values of γ_{js} are normally assumed to satisfy the constraints [11]:

$$\sum_{s=1}^{S} \gamma_{js} = 1 \;\; ; \;\; 0 \le \gamma_{js} \le 1. \qquad (2)$$

In HMMs, Gaussian mixture models are used to represent the output distribution of states. Equation (1) can be rewritten as:

$$b_j(o_t) = \prod_{s=1}^{S} [\sum_{m=1}^{M} C_{jsm} \mathrm{N}(o_{st};\mu_{jsm};\varphi_{jsm})]^{\gamma_{js}} \qquad (3)$$

where *M* is the number of mixture components in stream *s*, C_{jsm} is the weight of each mixture component of state *j* in each mixture of each stream and N*(O; μ, φ)* denotes a multivariate Gaussian of mean *μ* and covariance *φ.*

It is very important to choose proper exponents (or weights) since the performance of the system is significantly affected by the values of γ There has been a great deal of research on developing methods for optimizing the stream weights.

IV. Experimental Evaluation

A. Baseline System

In the Aurora project, whole-word HMMs were used to model the digits. Each word model consists of 16 states with three Gaussian mixtures per state. Two silence models were also considered. One of the silence models has relatively longer duration, modeling the pauses before and after the utterances with three states and six Gaussian mixtures per state. The other one is a single state HMM tied to the middle state of the first silence model, representing the short pauses between words [10].

In our experiment, the baseline system is defined over 39-dimensionel observation vectors, which consists of 12 cepstral and the log-energy coefficients, in addition the corresponding deltas and accelerations vectors. It is noted MFCC_E_D_A (39), and considered as the front-end by conventional DSR ETSI standard [3]. Training and recognition phases were carried out by the HMM-based toolkit HTK [15]. In some special cases HTK toolkit automatically divided the feature vector into multiple equally-weighted streams, in a way similar to the multi-stream paradigm. The idea of this separation is based on the lack of correlation between features. In the case of the baseline vector, we use three streams: one for the static coefficients plus an energy coefficient, the second and third are reserved, respectively, delta and acceleration coefficients with their delta and acceleration energy component.

B. Experimental Protocol

In case the front-end proposed in the framework of the multivariable, the 12 MFCCs coefficients and their first derivatives, without the energy component, are the first and second stream. The 10 coefficients MLSFs are taken as third stream. These multiple streams have equal weights. This new front-end will be noted by MFCC_D_MLSF (34), where 34 indicates its size. The MLSFs added to produce a multidimensional set of parameters, and replace the component accelerations and energies of conventional front-end. To assess the impact of weights γ_{js} of (1), we conducted another experiment on the same vector but with the use of different weights satisfying the (2). For example, the vector incorporated will be noted MFCC_D.8_MLSF.2 (34). It indicates a weighting of 0.8 for the first two flows and 0.2 for the third corresponding to MLSFs. On the other hand, we aim to optimize the use of these stream parameters by reducing the size of acoustic vectors while improving system robustness. For this, we apply a KLT on all flows constituting the vector. Also, MFCC_D.8_MLSF.2 (KLT_24) to indicate that KLT is applied to the vector of dimension 34, all three streams at different weights (0.8 for MFCCs and 0.2 for MLSFs), which we retain the first 24 components.

To further reduce noise, we propose to apply at the proposed acoustic frame, one floor of the Wiener filter. The estimate by the filter is individually made to short segments of the signal where two consecutive frames have a difference of time of 10 ms [9]. Also, the results presented in last line of the table 1 show that the features extraction from a denoising frame by the Wiener filter improves further the recognition rates for different SNR (especially those less than 5 dB), compared to the ETSI Mel-cepstral front-end and those using the MFCC-MLSF approach proposed. The constituted vector is noted by MFCC_D.8_MLSF^{dn}.2 (KLT_24). In our experiments, we opted for a KLT at class independent (CI-KLT) in which the transformation matrix is global and is determined for all classes, unlike the case of class dependent (CD-KLT) [12] where we use a transformation matrix for each acoustic model

Table I gives the results for GSM speech corrupted by different noise. Best results in terms of word recognition accuracy are edited in bold. For very low SNRs, when the SNR decreases less than 5dB, the use of MLSF front-end with 34-dimensionnal feature vector leads to a significant improvement in word recognition accuracy. We note that our approach is all the better, when the SNR decreases. The substituting of acceleration components and energy by the MLSFs in the baseline vector led to an improvement in rate recognition with a consequent dimension vector.

In addition, for a different weighting of flux used, there is a marked improvement from 10dB. At this level of SNR (10 dB), 20% contribution of MLSFs compared to MFCC improves the recognition rate significantly (up to 25%). On the other hand, we note that KLT applied to the new weighted vector, reduced in turn and optimizes the original space of parameters. KLT has led to better performance with fewer parameters. Under unfavourable conditions, the decomposition by KLT into subspaces works better compared to conventional front-end of ETSI.

Also, the results presented in last line of the table 1 (in bold for indicate the best score) show that the features extraction from a denoising frame by the Wiener filter improves further the recognition rates for different SNR (especially those less than 5 dB), compared to the conventional ETSI and those using the MFCC-MLSF approach proposed.

On the other hand, we note that KLT applied to the new weighted vector, reduced in turn and optimizes the original space of parameters. KLT has led to better performance with fewer parameters. Under unfavourable conditions, the decomposition by KLT into subspaces works better compared to conventional front-end of ETSI (Fig. 2).

V. Conclusion

In this paper, we investigate the performance of a new codec that could constitute an alternative to the present ETSI DSR-XAFE codec in severely degraded mobile environments. It is based on a multi-stream paradigm using a multivariable acoustic analysis Mel line spectral frequencies (MLSF) and MFCCs. The proposed system will be compatible with 3GPP and 3GPP2 standards respectively for both European (GSM) mobile and North American (CDMA) systems.

We conducted a further analysis of ETSI basic codec used from there recognition of speech distributed ETSI DSR basic. The results show that the MLSFs improve the performance of the DSR, compared to that of the basic DSR front-end of ETSI using MFCC alone. This improvement is especially important when we perform a pre-processing of KLT, more particularly for highly noisy environments. On the other hand, although the extraction the new features may add some level of complexity to the front-end process, the use of a 24-dimensional feature vector instead of 39-dimensional

TABLE I: RECOGNITION RATE (%) OF THE BASIC DSR SYSTEM AND THOSE USING THE MULTIVARIABLE ON AURORA DATABASE

Noise type	Multi-variable Vector	20db	15db	10db	5db	0db	-5db
Restaurant	MFCC-E-D-A (39)	89,99	76,24	54,77	31,01	10,96	3,47
	MFCC-D-MLSF (34)	81.89	75.93	62.02	38.62	18.70	9.64
	MFCC-D.8-MLSF.2 (34)	94.29	88.85	73.84	45.38	22.32	11.76
	MFCC-D.8-MLSF.2 (KLT_24)	92.97	89.25	75.87	50.14	23.06	12.54
	MFCC-D.8-MLSF^{dn}.2 (KLT_24)	**94.89**	**90.57**	**78.63**	**53.87**	**27.38**	**14.56**
Airport	MFCC-E-D-A (39)	90,64	77,01	53,86	30,33	14,41	8,23
	MFCC-D-MLSF (34)	74.77	66.03	51.00	33.46	17.63	9.07
	MFCC-D.8-MLSF.2 (34)	93.98	88.19	73.93	47.27	25.11	13.63
	MFCC-D.8-MLSF.2 (KLT_24)	92.34	88.34	75.66	49.45	26.59	14.94
	MFCC-D.8-MLSF^{dn}.2 (KLT_24)	**94.43**	**90.76**	**79.21**	**52.29**	**31.84**	**17.54**

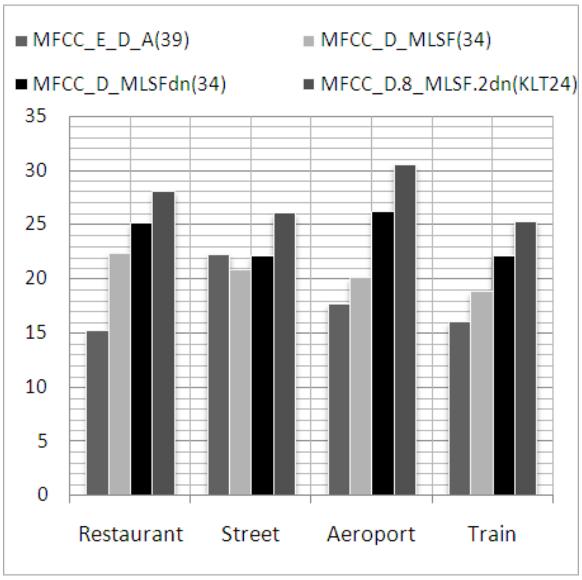

Fig. 2. Average recognition rate achieved with different types of test B noise for values 5, 0 and -5 dB SNR.

Feature vector will reduce the computing time and a storage capacity for the process performed on the main server. This work is being continued to assess the contribution of these new parameters in a noisy environment towards the auto-optimization of stream weight with respect to the noise source, speaker gender and phonetic contents of the speech.

REFERENCES

[1] D. Addou, S. Selouani, K. Kifaya, M. Boudraa, and B. Boudraa, "A noise-robust front-end for distributed speech recognition in mobile communications," *International journal of speech technologie*, Springer-verlag, vol.10, pp. 167-173, 2009.

[2] H. Bourland, S. Dupont, and C. Ris, *Multi-Stream Speech Recognition*. Tech. Rep, IDIAP, 1996.

[3] ETSI. Speech processing, transmission and quality aspects; distributed speech recognition; front-end feature extraction algorithm; Technical report of ETSI ES 201 108, 2003.

[4] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal Acoustic Society of America,* pp. 1738-1752, 1990.

[5] H. Hermansky and N. Morgan, "Rasta Processing of Speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578-589, 1994.

[6] K. Hermus and P. Wambacq, "Assessment of Signal Subspace Based Speech Enhancement for Noise Robust Speech Recognition," *IEEE Proceedings ICASSP*, 2004.

[7] I. T. Jolliffe. *Principal Component Analysis*. Second Edition. Springer, 2002.

[8] J. C. Junqua and J. P. Haton, *Robustness in Automatic Recognition*. Kluwer Academic Pub. AI series Springer-Verlag, New York, 1995.

[9] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC; 1 edition, 2007.

[10] D. Pearce and H. G. Hirsch, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," In *Proc. ICSLP*, vol. 4, pp. 29-32, 2000.

[11] R. Rose and P. Momayez, "Integration of multiple features sets for reducing ambiguity in automatic speech recognition," in *Proc. IEEE-ICASSP*, pp. 325-328, 2007.

[12] A. Sharma, K. K. Paliwal, and G. C. Onwubolu, Class-dependent PCA, MDC and LDA: A Combined Classifier for Pattern Classification Pattern Recognition, vol. 39-7, pp. 1215-1229, 2006

[13] H. Tolba, S. A. Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," in *Proc. of the ICASSP*, pp. 837-840, 2002.

[14] Y. Hu and P. C. Loizou, "A Subspace Approach for Enhancing Speech Corrupted by Colored Noise," *Signal Processing Letters IEEE*, vol. 9-7, pp. 204-206, 2002.

[15] S. J. Young, HTK version 3.4, Reference and User Manual. Cambridge University, 2006.

[16] T. Z. Hua, "Automatic Speech Recognition on Mobile Devices and over Communication Networks," Lindberg Børge (Eds.) Springer-Verlag, pp. 115, 2008.