Distributing Recognition in Computational Paralinguistics

Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller, *Member, IEEE*

Abstract—In this paper, we propose and evaluate a distributed system for multiple Computational Paralinguistics tasks in a client-server architecture. The client side deals with feature extraction, compression, and bit-stream formatting, while the server side performs the reverse process, plus model training, and classification. The proposed architecture favors large-scale data collection and continuous model updating, personal information protection, and transmission bandwidth optimization. In order to preliminarily investigate the feasibility and reliability of the proposed system, we focus on the trade-off between transmission bandwidth and recognition accuracy. We conduct large-scale evaluations of some key functions, namely, feature compression/decompression, model training and classification, on five common paralinguistic tasks related to emotion, intoxication, pathology, age and gender. We show that, for most tasks, with compression ratios up to 40 (bandwidth savings up to 97.5 percent), the recognition accuracies are very close to the baselines. Our results encourage future exploitation of the system proposed in this paper, and demonstrate that we are not far from the creation of robust distributed multi-task paralinguistic recognition systems which can be applied to a myriad of everyday life scenarios.

Index Terms—Computational paralinguistics, distributed recognition system, split vector quantization, emotion

1 Introduction

In recent years, computational paralinguistics has attracted the attention of speech and language processing researchers due to its prominent potential for practical applications in everyday life scenarios [1], [2]. For instance, it can support the interpretation of people's stable and transitory states, such as their intentions, emotional and mood states, confidence and stress levels, physical condition, age, gender, personality traits, amongst many others. From a computer science perspective, such information is crucial for the improvement of machine mediated human-human interactions as well as human-machine interactions. An example are call centers, as it can support, for instance, the detection of negative emotional states in customers which in turn can inform the development of new strategies to ameliorate the interactions and service provided.

1.1 Embedded versus Client-Server-Based Recognition Systems

Most state-of-the-art academic research in this area focuses on statically embedded recognition [1], [2], [3], [4]. Such systems have a good degree of flexibility since they can be used without Internet access and therefore applied in a wide range of practical scenarios. Nonetheless, at present, this advantage is becoming less relevant. On the one hand, because current data-driven pattern recognition systems largely benefit from processing large amounts of data for training and continuous development, which requires data transmission for the integration of data from multiple users (as well as vast storage and computational resources for training). Furthermore, sophisticated computational paralinguistic systems may require advanced computational models whose implementation is not feasible in users' devices [5]. On the other hand, because internet access is now ubiquitous on account of the advent of far-ranging coverage and high-speed wireless networks such as 3G, 4G, and wireless LAN, and the breakout of mobile electronic devices like smartphones, laptops, and tablets.

One possible solution to this problem is to recur to client-server computing [6]. On the client side, the normal consumer devices with restricted computing ability can perform basic computational tasks, while, on the server side, super computers or computing centers can deal with the most expensive computational tasks. In the context of computational paralinguistics, the client is responsible for collecting realistic data (i.e., voice recordings in natural occurring scenarios) which is then sent to the server. In the server, the computational resources can be employed to integrate the data from the various clients, build (and continuously improve) the target paralinguistic system(s), classify the data received from the various clients for the task at hand, and feedback the final results to the various clients.

Such a solution has several advantages for the future development and application of paralinguistic recognition systems in real-life scenarios. First, it can overcome one of the most important limitations for the development of robust paralinguistic recognition tasks—*data scarcity*. In contrast to other pattern recognition areas like automatic

- *Z. Zhang and J. Deng are with the Machine Intelligence and Signal Processing Group, Technische Universität München, 80333, Munich, Germany. E-mail: {zixing.zhang, jun.deng}@tum.de.*
- *E. Coutinho is with the Machine Intelligence and Signal Processing Group, Technische Universität München, 80333, Munich, Germany and the School of Music, University of Liverpool, United Kingdom. E-mail: e.coutinho@tum.de.*
- *B. Schuller is with the Machine Intelligence and Signal Processing Group, Technische Universität München, 80333, Munich, Germany and the Department of Computing, Imperial College London, United Kingdom. E-mail: schuller@tum.de.*

speech recognition, the data for computational paralinguistics is quite scarce, which constitutes a major concern in this field that needs to be addressed [1], [2], [7]. To this goal, client-server-based systems have the potential to allow the collection of large amounts of labeled and unlabeled realistic data from thousands of users in real-life scenarios, which can be exploited for training models and enhancing their performance using machine learning techniques like semi-supervised learning [8], co-training [7], active learning [9], or even advanced crowd sourcing [10]. Such data enrichment and optimization techniques are of crucial importance to continuously enhance the systems' robustness [8], without the need of users to exchange data on the client side. Second, it can accelerate the improvement of paralinguistic recognition systems since having the data processed on the server side computer scientists can continuously develop and apply more effective techniques (e.g., bidirectional long short-term memory [11] or cumulative evidence [12]) and combine various data sources to boost the systems' performance and robustness. Moreover, user profiles can be stored in the server to support long-term analysis and improve user-specific models. Third, on the client side, the requirements of computing power, the conditions of operating systems and hardware configurations are greatly relaxed, therefore making it possible to spread the use of paralinguistic analysis to a wide range of personal mobile or fixed devices.

1.2 Network versus Distributed Recognition Systems

Concerning the location where the feature extraction takes place, client-server architectures for computational paralinguistics can be categorized into two classes: network recognition systems and distributed recognition systems [6]. The former uses conventional speech coding for the transmission of speech from a client device to a server where decoding and feature extraction are undertaken. The latter implies that the feature extraction stage is processed on the client side, but the recognition is made on the server.

One of the major advantages of adopting a network recognition approach is that it is not necessary to develop a completely new system for paralinguistic recognition tasks. Indeed, numerous commercial applications already implement speech coding, and so, without the need to change the applications on existing devices and networks, we can simply use preexisting recognition models on the server side to process the encoded speech signals. Moreover, it shares all the advantages of server-based systems in terms of system maintenance, update, and device requirements [6]. Nonetheless, as it will be discussed in the next section (Section 1.3), network recognition systems posit various challenges related to privacy and transmission bandwidth limitations.

In distributed pattern recognition systems, instead, the feature extraction process occurs on the client side, where a representation of the speech signal with a lower dimensionality and redundancy can be obtained and optimized for transmission. Such systems have been adopted in various applications, being some of the most impressive and successful ones developed in the context of speech recognition, where both theoretical (e.g., packet loss via transmission [13], feature compression techniques [14], and noise robustness [15]) and applied (e.g., Google search engines and Apple's Siri [16]) research has been conducted. In other fields, distributed pattern recognition has also been applied, for instance, to the recognition of human actions through the use of wearable motion sensor networks [17], nature elements (such as trees or weeds) or faces [18].

In relation to computational paralinguistics, if distributed computing can be demonstrated to be feasible and reliable, some of the current limitations preventing recognition systems to be applied to large-scale realistic applications can be greatly mitigated. More importantly, this would be beneficial to a variety of areas, such as, remote medicine treatment, remote conferences or negotiations, remote education, and even advanced driver assistance systems, where paralinguistic recognition systems have manifold applications.

In this paper, we propose a distributed recognition framework for paralinguistic tasks inspired by the standardization work of distributed speech recognition done by the Aurora group from the European telecommunications standards institute (ETSI) [19]. Compared to our previous work described in [20], where we only targeted the recognition of emotional states, here we provide a detailed description of a distributed recognition system and evaluate its application to large-scale and realistic tasks pertaining to three different time-scales of paralinguistic phenomena (following [2] and [1] taxonomy for the categorization of paralinguistic phenomena in the context of computer science): 1) short-term states, e.g., emotions and emotion-related states or affects (stress, confidence, interest, frustration, etc.); 2) medium-term phenomena, like health state, intoxication, sleepiness, mood (depression), friendship; and 3) long-term traits, such as biological aspects (e.g., age and gender), personality-related features (likability), and social background (culture, race, status, etc.). In particular, we will focus on feature (de)compression and paralinguistic information recognition systems with the aim of dealing in efficient ways with transmission bandwidth limitations and warranting users' privacy, which are core aspects for the future application of such framework.

1.3 Privacy and Transmission Bandwidth Limitations

One major concern in paralinguistic recognition is *security* since the privacy of the speakers has to be guaranteed. This is particularly important in real-life contexts where personal information and sensitive data may be collected. Paralinguistic information is indeed of highly private nature as it can contain, for instance, emotional statements, information about alcohol intoxication, tiredness, etc. [21].

The transmission of raw coded speech is a common approach in client-server architectures. The speech data are normally coded by protocols like G.711, G.726, and AMR-WB [22]. However, these coding protocols target a faithful recovery of speech for better communication quality, which could lead to exposure of user personal information. As mentioned earlier, a possible alternative is to perform feature extraction directly in the client and transmit low-level-descriptors (LLDs) [19], therefore preventing direct access to users' speech. Unfortunately, previous work has shown that it is feasible to reconstruct the audio from static feature

TABLE 1
Turn Duration (Average, Minimum, Maximum, and Standard Deviation) and Required Transmission Bandwidth for Three Transmission Strategies—**raw** Coded Speech, **LLDs**, and **stat**istical Feature Set—and All Corpora Used in This Paper (AEC, ALC, NCSC, and Agen; A Detailed Description of the Databases and Respective Acronyms is Given in Section 4.1)

	uttr. length (s)				bandwidth (kb/s)		
Corpus	avg	min	max	std	raw	LLDs	stat
AEC	1.7	0.1	24.5	0.8	$16 \sim 40$	51.2	7.3
ALC	11.4	1.5	61.8	14.2	$16 \sim 40$	188.8	12.3
NCSC	3.1	0.9	21.2	1.8	$16 \sim 40$	204.8	62.4
Agen	2.6	0.3	11.3	1.2	$16 \sim 40$	92.8	5.5

vectors like Mel frequency cepstral coefficients (MFCCs) and pitch (e.g., [23], [24]) which is why they are used for speech recognition and speaker recognition. This once more generates important privacy-related concerns.

For our system, we propose to generate and transmit statistical feature vectors obtained by applying functionals over LLDs for each utterance. The procedure of generating such feature vectors is irreversible, and therefore they avoid the reconstruction of the speech signal and permit to overcome the issue of users' privacy violation. Because of this irreversibility, the speakers' speech content is fully protected, which is significantly important since the speech content is widely admitted as the most important personal information. Moreover, even though access to statistical features could be used to infer private information (e.g., age or gender of speakers), that would be only possible by having access to the computational models stored in the server (something that is extremely unlikely). Additionally, in the context of state-of-the-art computational paralinguistic research, statistical features are nowadays a well-accepted standard for extracting relevant information from speech (e.g., [1], [2], [25]).

Another major concern of network-based systems with relevance to our work is *transmission bandwidth*. Let us consider as an example the official databases of the INTERSPEECH 2009-2012 Challenges [4], [26], [27], [28], which will be used in this paper to evaluate our framework. We calculated the bandwidth necessary for each of the coding strategies mentioned above (raw coded speech, LLDs, and statistical features) for various turn durations (the ITU-G.726 protocol was considered for coding raw speech, and single precision floating point—32 bit—was used for LLDs and statistical feature sets). In the case of statistical features, given that the vector dimensionality is always the same per turn (and so the transmission bandwidth will vary as a function of turn duration), the bandwidth size was calculated for the average turn duration in each data set. As it can be observed from Table 1, with the exception of the pathology task, the statistical feature set requires less bandwidth than the remaining coding strategies.

Statistical feature sets seem to satisfy privacy concerns and require less bandwidth than raw coded speech and LLDs transmission. However, such an approach still requires a large bandwidth if we consider a target scenario involving a large number of users/devices. Therefore, it is necessary to reduce further the dimensionality of the feature space (while taking into account the recognition performance). There are at least two general solutions which can be considered to deal with a limited transmission bandwidth: feature selection and feature compression. A feature selection strategy takes into account the features' relevance, irrelevance and abundance, and aims at selecting a subset that can predict the output with an accuracy which is comparable to the performance of the complete feature set. Typical methods achieving this goal include wrappers, filters, and embedded routines [29], [30]. Some algorithms, such as minimum redundancy maximum relevance [31] and random subset feature selection [29], are now well-developed, and have been successfully applied to paralinguistic tasks [32], [33]. In this paper, nevertheless, we do not explore the use of feature selection algorithms, nor the merit of individual features in the original space, which has repeatedly been explored in the literature (cf. e.g., [34]). Instead, we employ feature sets which have been previously optimized for the various paralinguistic tasks used in this work (cf. Section 4.2). There are two main reasons for this. First, we intend to focus only on the essential components of the distributed system. Feature selection techniques can easily be integrated into the system as a 'plug-in' [32]. Second, in order to directly and fairly compare the performance of the distributed system with the baseline performances of embedded systems the same features sets should be used.

Feature compression generally refers to methods that transform a high dimensional feature space into a lower dimensional one. Typical dimensionality reduction methods include principle component analysis (PCA) [35] and linear discriminant analysis (LDA) [36], and have been implemented, for instance, in distributed face recognition [37], speaker identification [38], and speech recognition [39]. Another family of methods for (lossy) compression is vector quantization (VQ) [40], which has been very popular in a variety of research fields such as speech coding [41], image and video compression [42], and various pattern recognition applications (e.g., face detection [43], texture classification [44]). There are many variations of VQ proposed in the literature, such as, distance-based VQ [45], histogram-based quantization (HQ) [46], lattice VQ, and address VQ [47]. Concerning the work presented in this paper, we opted for a particular VQ compression algorithm known as split vector quantization (SVQ) [48]. The main reasons for choosing SVQ are: i) the assignment of prototype numbers from a codebook eliminates any direct feature information from the user, thus ensuring privacy [19]; ii) SVQ is the officially recommended method by the ETSI standards [19] for distributed speech recognition; and iii) it is a well established and efficient compression technique [40], [41], [47].

The remainder of this paper is organized as follows. Section 2 describes a unified distributed recognition system for paralinguistic tasks. In Section 3 we describe the SVQ feature compression method, and in Section 4 we introduce the four corpora covering short-term, medium-term, and long-term paralinguistic tasks for classification used in this paper. In Section 5 we evaluate the impact of feature independence for SVQ on the proposed system, and present the results for a large-scale experiment on five paralinguistic tasks. Finally, in Section 6, we deliver our conclusions and

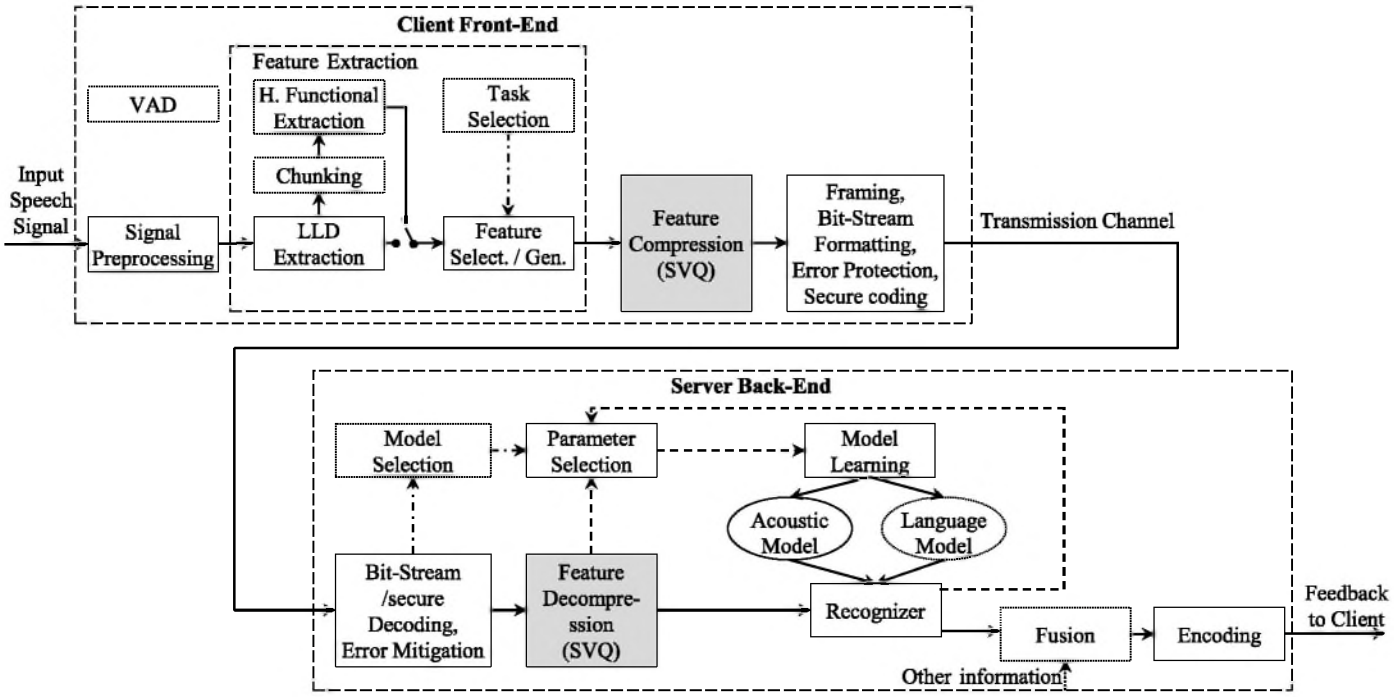


Fig. 1. Proposed framework for distributed paralinguistic recognition system. *Dotted boxes* indicate optional components. *Dashed lines* show steps carried out only during system training or adaptation phases. *Dash dotted* lines indicate steps carried out when multiple recognition tasks need to be processed at the same time or separately.

discuss the strengths, limitations, and implications of the work presented in this paper.

2 Overview of the Distributed Recognition System for Paralinguistic Tasks

Fig. 1 illustrates the framework of the distributed paralinguistic recognition system proposed in this paper. It consists of two main modules: client and server.

2.1 Client Module

The client module consists of several sequential processing stages that aim at maintaining as much information as possible about the speech signals while using the less transmission bandwidth possible, and dealing with transmission distortion.

First, a voice activity detection (VAD) algorithm is used for detecting speech signals and dropping the non-speech frames. If speech is detected, the signals are then delivered to the following processing components. Given that the incoming speech signals are always distorted by various noise sources, such as additive noise of multiple speakers, environmental, and recording noise, as well as convolutional noise like reverberation, the goal of the speech preprocessing stage ("Signal Processing" block) is to enhance the incoming speech signals and filtering out undesired signals. Common techniques to deal with these problems include adaptive filters [49], spectral normalization and subtraction [50], non-negative matrix factorization [49], and beamforming [51].

Following the "Signal Processing" stage, the denoised and enhanced signals are sent to the feature extraction module. Considering the advantages highlighted in Section 1.3 regarding privacy and bandwidth requirements, we propose the encoding of the speech signals using statistical features computed on LLDs rather than raw coded speech or simply LLDs. LLDs were computed at approximately 100 frames per second with typical window sizes of 10-30 ms. The windowing functions used for extraction of LLDs in the time or time-frequency domain are typically smooth (Hamming or Hann) or rectangular [2]. Then, the LLD sequence is divided (chunked) into 'super-segmental' turns, and functionals over LLDs are applied to each turn ("Chuncking" block). The turns can be a fixed number of frames, syllables, words, acoustic chunks, sub-turns, or complete turns [52]. In our framework, the selection of turns depends on the requirements of the specific recognition tasks (e.g., the emotion-related information often involves transient speech, while the gender information can cover the whole speech track). For the experimental work presented in this paper, we recur to the LLD set used in the INTERSPEECH 2009-2012 Challenges [4], [26], [27], [28], whose dimensionality per frame ranges from 16×2 to 60×2 (the exact number of features depends on the recognition task) if the derivatives are also adopted. By applying functionals to each trunk, the final transmission bandwidth requirements lies between 5.5 kb/s and 62.4 kb/s (cf. Table 1).

As discussed earlier, in order to further reduce the required transmission bandwidth a feature compression stage is useful (a detailed description will be provided in Section 3). Following feature compression, and before the compressed features enter into the physical transmission channel, framing, bit-stream formatting, error protection, and secure coding algorithms are necessary in order to meet the transmission requirements (e.g., IP routing, clock recovery), prevent channel distortion (e.g., channel noise, packet loss), and guarantee information security.

Given the need of dealing with various tasks simultaneously, and the fact that information required for a specific

task may also be relevant to other tasks, task selection algorithms are definitely important. A common way is to allow the client to perform task selection which in turn would determine the feature set chosen on the client side and the model selected on the server side. If no specific tasks are predefined, however, computational auditory scene analysis (CASA) could be used to automatically analyze the circumstances of speech recording (e.g., driving, cocktail party, home, street) and determine the possible tasks [53]. It is nevertheless out of the scope of this paper to introduce automatic task selection.

2.2 Server Module

Turning now to the server module, we start by including bit-stream and secure decoding as well as error mitigation in order to recover the transmitted signal and convert it back to the compressed feature set. Then, the feature set is decompressed into its corresponding higher dimensional set ("Feature Decompression (SVQ)" stage) by using the codebook generated by the server (the codebook allows the translation from the compressed feature space to the original one, and vice-versa). Next, the decompressed feature set is delivered to the recognizer for classification, which in turn outputs discrete labels which are associated with the particular recognition task (e.g., positive/negative arousal) or regression values when the output is a continuous quantity (e.g., speaker's height or age). The classification/regression results from the acoustic and language models can also be integrated with other information such as facial expressions, motion patterns, among others. Finally, the relevant feedback information is encoded and transmitted to the client through the network.

With respect to the training and adaptation of server-side models, several methods can be effective: supervised learning, semi-supervised learning, combinations of both, among others. For instance, a particular model in the server back-end can use annotated speech received from the various clients for model training. Other possibilities include the use of unlabeled data to improve the models, by applying, for instance, Co-Training techniques [7]. Compared to traditional paralinguistic recognition systems, it is easier and cheaper to collect large amounts of data from different contexts in the realistic world, giving rise to the ability of training a more robust model by a powerful server. In this process, parameter optimization is also required to suit different classification algorithms, e.g., the kinds of kernel and complexity constants of support vector machines (SVMs), neural networks topology, etc.

3 Feature (De)Compression

SVQ algorithms split the high dimensional feature vectors into several subvectors which automatically group the original feature set through some sort of clustering algorithm (e.g., k-means). Each subvector is then represented by the centroid of each group.

Fig. 2 shows a diagram depicting the SVQ algorithm. The encoding scheme firstly partitions the whole r-dimensional feature vector $F = [f_1, f_2, \ldots, f_r]^T$ into P subvectors, $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_P]^T$, each of which has k dimensions. Thus, r is equal to the sum of dimensions of each subvector,

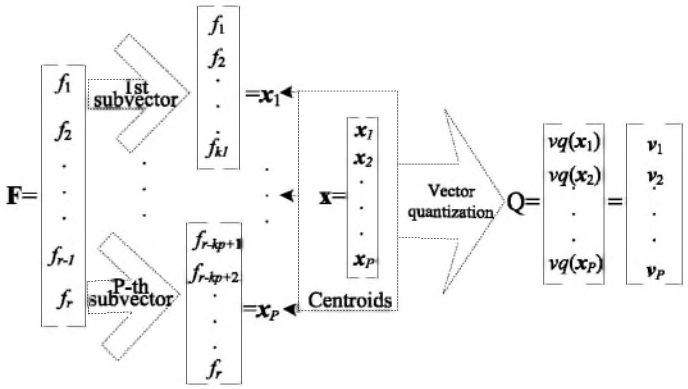


Fig. 2. Diagram of split vector quantization algorithm.

$r = k_1 + k_2 + \cdots + k_P$. In the particular case of having the same number of dimensions in each subvector, then $r = k \times P$. Following, each subvector is quantized using a different VQ codebook, $\boldsymbol{Q} = vq(\boldsymbol{X}) = [vq(\boldsymbol{x}_1), vq(\boldsymbol{x}_2), \ldots, vq(\boldsymbol{x}_P)]^T = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_P]^T$, where $\boldsymbol{v}_i \in \boldsymbol{C}_i$. Note that the codebook ($\boldsymbol{C}_i$) pertaining to a particular subvector can be different from that of other codebooks, not only in the clustering space but also in size.

In our implementation, we used a k-means algorithm for clustering. That is, each observation belongs to the closest quantization centroid which is found by using a weighted euclidean distance to determine the index:

$$\boldsymbol{d}_i^j = \boldsymbol{x}_i - \boldsymbol{v}_i^j, \quad i = 1, \ldots, P; j = 1, \ldots, N_i, \tag{1}$$

$$idx_i = \underset{1 \le j \le (N_i)}{\arg\min} \left(\boldsymbol{d}_i^j\right) W_i \left(\boldsymbol{d}_i^j\right), \tag{2}$$

where $\boldsymbol{v}_i^j$ is the jth codevector in the codebook C_i, $\boldsymbol{d}_i^j$ denotes the Euclidean distance between subvector $\boldsymbol{x}_i$ and codevector $\boldsymbol{v}_i^j$, N_i is the size of the codebook, W_i is the weight matrix, e.g., identity matrix, to be applied to the codebook C_i, and the idx_i denotes the codevector index chosen to represent the vector $\boldsymbol{x}_i$.

The final set of quantized vectors, $[idx_1, idx_2, \ldots, idx_P]^T$, is used to represent the corresponding speech chunk, and is transmitted to the server back-end. On the server back-end, the SVQ process is reversed by using the same codebook used in the front-end for each subvector:

$$\hat{\boldsymbol{x}}_i = \boldsymbol{v}_i^{idx_i}, \tag{3}$$

where $\hat{\boldsymbol{x}}_i$ denotes the estimate of $\boldsymbol{x}_i$. Then, we unify all estimated subsets of features into a single vector, $\hat{F} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \ldots, \hat{\boldsymbol{x}}_P]^T$.

Finally, it is important to mention that there are various aspects that need to be taken into consideration when using SVQ as they can impact the performance of the various recognition task. Those will be dealt with in Sections 5.1 and 5.2.

4 Databases and Acoustic Features

4.1 Databases

In this paper we recur to four frequently used paralinguistic databases to test our system: the FAU Aibo emotion corpus (AEC), the alcohol language corpus (ALC), the NKI CCRT speech corpus (NCSC), and the Agender database. The

TABLE 2
Overview of Selected Corpora for Emotion (AEC), Intoxication (ALC), Pathology (NCSC), Age and Gender (Agen) Recognition Tasks

Corpus	LA	TY	S #	TT[H]	TC[s]	INST #	Hz
AEC	DE	S	51	8.9	1.7	18,216	16 k
ALC	DE	P	162	43.8	11.4	12,360	16 k
NCSC	NL	P	55	2.0	3.1	2,386	16 k
Agen[1]	DE	P	770	35.9	2.6	53,074	8 k

Languages (LA): German (DE) and Dutch (NL); speech types (TY): spontaneous (S) and promoted (P); number of subjects (S) and instances (INST); total speech time (TT) and average speech time per chunk (TC); recording rate (Hz).
[1]*Test labels of Agender are not freely available. Thus, only its partitions of train and develop are used in our experiments.*

tasks associated with the four corpora cover a variety of time-relations of paralinguistic groups from the short-term (emotion), medium-term (intoxication and pathology), and long-term (age and gender) phenomena. Further details on the four corpora are shown in Table 2. Table 3 shows the speaker-independent partition of instances. In what follows we briefly describe each of the four databases.

4.1.1 Emotion: FAU Aibo Emotion Corpus

The FAU Aibo emotion corpus [25] is the official corpus of the INTERSPEECH 2009 emotion challenge (EC) [26]. This corpus contains audio recordings of German-speaking children interacting with Sony's pet robot Aibo [25]. For the construction of this database, children were led to believe that the Aibo was responding to their commands by producing a series of fixed and predetermined behaviors. Nevertheless, the Aibo robot did sometimes disobey to the children's commands, which provoked various types of emotional reactions. The recordings include speech samples from 51 children (30 females) with ages ranging from 10 to 13 years from two different German schools, MONT and OHM. The various recordings were labeled using two cover classes: one consisting of **NEG**ative emotion labels (*angry*, *touchy*, *reprimanding*, *emphatic*), and the other (**IDL**e) consisting of non-negative states (for more information about the database development and data processing please refer to [26]).

4.1.2 Intoxication: Alcohol Language Corpus

The alcohol language corpus [54] is the official corpus of the Intoxication Sub-Challenge from the speaker state challenge (SSC) at INTERSPEECH 2011 [27]. The database includes speech recordings of various people with ages ranging from 21 to 75 years old, either sober or with blood alcohol concentrations (BACs) ranging from 0.28 to 1.75 per mill. Three different speech recording conditions were conducted: read speech, spontaneous speech, and command & control. For our experiments, the recordings from speakers with BAC $\leq$ 0.5 per mill were labeled as non-alcoholized (**NAL**). All other instances were labeled as alcoholized (**AL**).

4.1.3 Pathology: NKI CCRT Speech Corpus

The "NKI CCRT speech corpus" [55] is the official corpus of the Pathology Sub-Challenge of the INTERSPEECH 2012 speaker trait challenge (STC) [28]. The database was created at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute and consists of speech recordings from 55 speakers (10 female; mean age of 57 y.o.) before and after chemo-radiation treatments (CCRT). All speakers read a text in the Dutch language with an emotionally neutral content. Thirteen speech pathologists evaluated the speech recordings in an online experiment on an intelligibility scale ranging from 1 to 7. Finally, an evaluator weighted estimator was used to compute and discretize the ratings into binary classes—intelligible (**I**) and non-intelligible (**NI**)—using the median of the ratings distribution.

4.1.4 Age and Gender: Agender

The Agender database [56] is the official corpus of the INTERSPEECH 2010 paralinguistic challenge (PC) Age and Gender Sub-Challenges [4]. This database was collected by a commercial company with the aim of identifying people of specific targeted ages and genders. The participants were asked six times to call an automated Interactive Voice Response system and to repeat various German utterances or to produce free speech content. The calls were made through a mobile phone in various recording environments, and in different days and times so as to ensure more variation in the voices of each speaker. In the Challenge task, four classification classes were used for age—**C**hildren, **Y**oung, **A**dult, and **S**enior—and three for gender—**C**hildren (**X**), **M**ale, and **F**emale. Additionally, here we also consider seven new classes which are generated by combining the various age and gender classes. Hereinafter, we will refer to this classification task as "Age + Gender" (for more details please refer to [4]).

4.2 Acoustic Features

As introduced in Section 2, the feature set was computed by applying functionals over LLDs per trunk and corresponding first and/or second order delta regression coefficients.

TABLE 3
Instances Distribution per Partition (Train, Develop, or Test) for Four Paralinguistic Corpora—AEC, ALC, NCSC, and Agender

	AEC		ALC		NCSC		Agender: Age				Agender: Gender		
	NEG	IDL	NAL	AL	I	NI	C	Y	A	S	X	M	F
Train	3,358	6,601	3,750	1,650	384	517	4,406	8,657	8,990	10,473	4,406	13,985	14,135
Develop			2,790	1,170	341	405	2,396	4,892	5,873	7,387	2,396	8,508	9,644
Test	2,465	5,792	1,620	1,380	475	264							
Σ	5,823	12,393	8,160	4,200	1,200	1,186	6,802	13,549	14,863	17,860	6,802	22,493	23,779

NEG: negative; IDL: idle; (N)AL: (non-)intoxicated; (N)I: (non-)intelligible; C/Y/A/S: children/young/adult/senior; X/M/F: children/male/female. No development set is defined on the AEC.

TABLE 4
Features Used for Five Paralinguistic Tasks. (Int.: Intoxication)

# Features	Emotion	Int.	Pathology	Age/Gender
LLDs	16	59	64	29
Functionals	12	39	61	8
Total	384	4,368	6,125	450

For our experiments, we use the same feature sets used in the INTERSPEECH Challenge 2009-2012 for emotion, intoxication, pathology, age and gender tasks. All features were extracted using the openSMILE toolkit [57]. Table 4 gives a detailed overview of the features used.

The feature sets sizes of the various tasks range from 384 (emotion recognition) to 6,125 (pathology task). The acoustic LLDs contain: energy-related features, such as frame energy, frame intensity/loudness, and zero-crossing rate; spectral-related features, such as cepstral coefficients (MFCCs, etc.), line prediction cepstral coefficients, and line spectral pairs; voice-related features, like perturbation (jitter, shimmer, etc.), harmonicity (harmonics-to-noise ratio, noise-to-harmonics ratio, etc.), fundamental frequency, and probability of voicing; and linguistic features, e.g., length of words, fragments, repetitions. In relation to the functionals used, they include extreme values and position (maximum, minimum, etc.), mean (arithmetic, quadratic, etc.), moments (standard deviation, variance, skewness, kurtosis, etc.), percentiles and percentiles range, regression (linear and quadratic approximation, etc.), centroid, peaks (number, distance, etc.), segments (number, duration, etc.), spectral (Discrete Cosine Transformation coefficients, etc.) and temporal (durations, positions, etc.) parameters. For full details on the feature sets please refer to [26], [4], [27], and [28].

5 Experiments and Results

In our classification experiments, we adopt linear SVMs trained with the sequential minimal optimization algorithm as implemented in the Weka toolkit [58], in line with the INTERSPEECH 2009-2012 EC [26], PC [4], SSC [27], and STC [28] challenges. Furthermore, we follow the classifier set-ups for the five Sub-Challenges. The complexity constants were optimized on the development set or through cross-validation of the training set (depending on the task). The resulting values were 0.05, 0.01, 0.001, 0.05, and 0.05 for emotion, intoxication, pathology, age, and gender tasks, respectively. Furthermore, as in the challenge, to alleviate the influence of instance imbalance, we implemented instance upsampling before any learning process, which produces a random subsample of the dataset belonging to sparse categories with-/out replacement.

In relation to the classification performance evaluation, we recur to the unweighted average recall (UAR; the sum of the recalls per class divided by the number of classes), which is the performance measure used in the 2009-2012 INTERSPEECH Challenges. In our experiments on ALC and NCSC tasks, the training and development sets are combined for training, and the test sets are used for testing. For the AEC task, given that there is no development set, only the training set is used for training (and the test set for testing). In relation to the Agender task, the development set is used for testing given that there is no test set. The UAR baselines for the binary classification on the emotion, intoxication, and pathology classification tasks are 67.6, 66.0, 69.0 percent, respectively. The baseline for the three-class gender classification is 76.0 percent, and the baseline for the four-class age classification is 45.7 percent. It should be pointed out that the baselines obtained in this paper are different from those reported in the 2009-2012 INTERSPEECH due to the use of a different Weka version.

For the sake of simplicity, in each paralinguistic task we split the whole feature vector (r dimensions) into multiple subvectors with the same dimensionality k (note that the last subvector dimensionality may be smaller than k and equal to r mod k). We also adopted the same codebook size N ($N = 2^L$, where L is codevector length) for all subvectors. In this case, the transmission bandwidth B_w for such compressed features is

$$B_w = \left(\left\lceil \frac{r}{k} \right\rceil \cdot L\right) \Big/ T. \tag{4}$$

Hence, its corresponding feature compression rate R for a transmission bandwidth requirement $B_{w/o}$ (no feature compression) is calculated by the equation

$$R = \frac{B_{w/o}}{B_w} = \frac{(32 \cdot r)/T}{(\lceil \frac{r}{k} \rceil \cdot L)/T} \cong 32 \cdot \frac{k}{L}, \tag{5}$$

where L is the length of codevector, T is the average length of a chunk, and assuming a single-precision floating point for the transmission of uncompressed data (32 bits). Obviously, the feature compression rate R is in direct proportion to the subvector dimension k and in inverse proportion to the codevector length L.

5.1 Influence of Attributes Independence

As discussed in Section 3, a central issue of SVQ is the splitting of the whole feature set into multiple subvectors in an effective way. The most important factor is arguably the cross correlation of attributes in the feature domain. A simple method to deal with this issue is to adopt a splitting strategy based on the types of LLDs, that is, the statistical features belonging to the same LLD are grouped into one subvector. In order to test this method, we compared the performance of this strategy with the performance achieved using a random clustering of the features on the five paralinguistic tasks. The dimensionality of all subvectors was set to the same value in each task—$k = 12, 37, 35, 8, 8$ for emotion, intoxication, pathology, age, and gender recognition, respectively. Given that the number of functionals over each LLD within each task may be different, we defined the dimensionality of the subvectors for each task as the maximum number of functionals over all LLDs. Table 5 shows the results obtained for the various tasks.

As it can be seen in Table 5, the performance achieved through LLD-based vector splitting strategy is always better than the strategy that used a random splitting strategy. This improvement is evident for all codebook sizes and across all tasks, and lies in the range of $1 \sim 3$ percent (absolute UAR). Results also show that the improvement delivered by the LLD-based splitting strategy over the random one is more noticeable for the tasks with larger features spaces,

TABLE 5
Performance Comparison for Five Paralinguistic Tasks Using Two Types of Vector Splitting Strategies: LLD-Based (D) and Random (R)

UAR	BL	$k=$	$N=128$		$N=256$		$N=512$	
[%]			D	R	D	R	D	R
Emo	**67.6**	12	66.1	65.3	66.7	65.8	67.4	66.8
Int	**66.0**	37	61.4	59.7	63.2	60.4	61.9	60.7
Path	**69.0**	35	69.0	66.6	68.3	66.8	69.1	66.2
Age	**45.7**	8	44.5	43.6	44.6	43.8	44.9	44.0
Gen	**76.0**	8	75.0	73.9	75.3	74.1	75.2	74.2

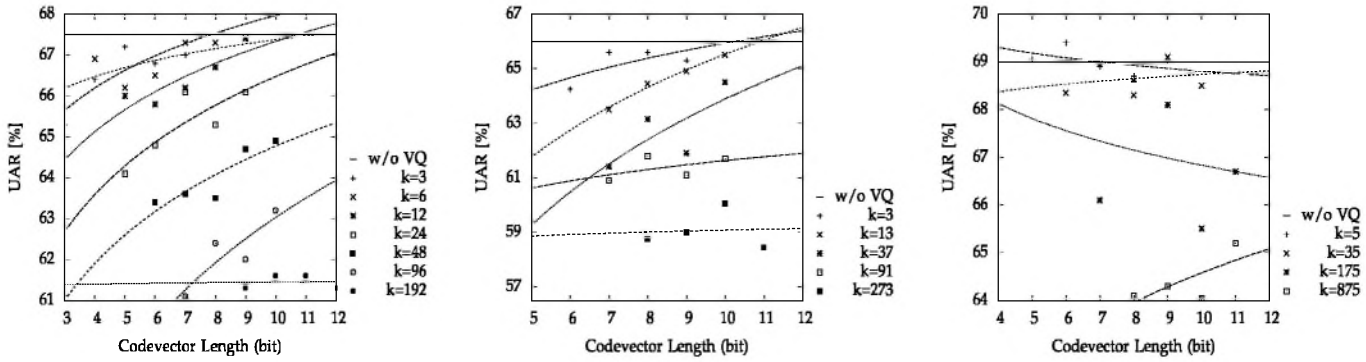
*BL: baseline; k: dimension of subvector; N: codebook size for each subvector. **Emotion**, **Intoxication**, **Pathology**, **Gender**.*

i.e., Intoxication (absolute average improvement of 1.9 percent) and Pathology (absolute average improvement of 2.3 percent). The absolute average improvement on the Emotion, Age and Gender tasks is less pronounced: 0.8, 0.8, and 1.1 percent, respectively.

5.2 Distributed Paralinguistic Tasks Classification

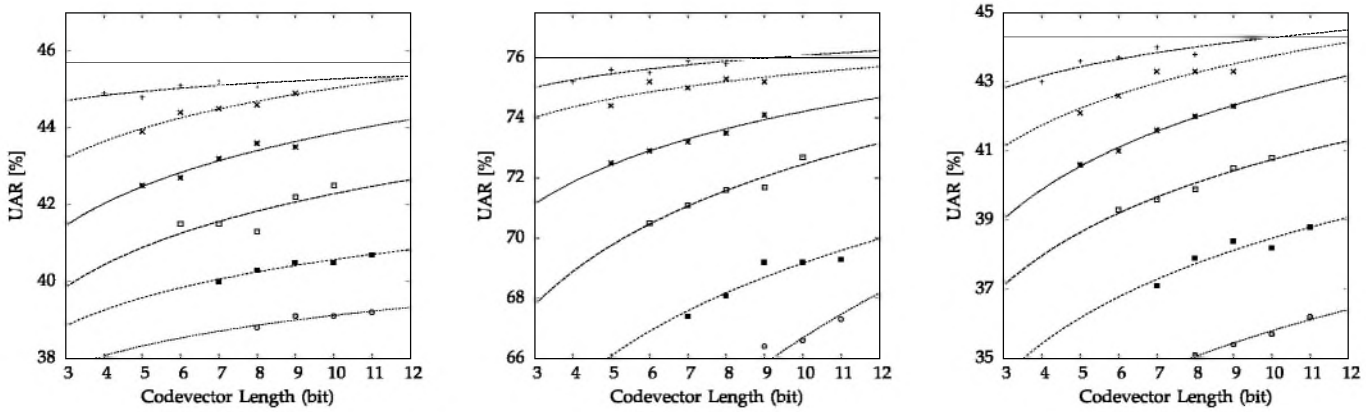
In the context of distributed speech recognition, the feature set typically comprises 14 features, and adjacent features are grouped into pairs [19]. This is quite different from distributed paralinguistic tasks, where the feature spaces are much larger (cf. Table 4). Therefore, grouping features into pairs would lead to a very large number of subvectors and low compression rates, which is not ideal given the bandwidth limitations. In order to investigate the influence of the dimensionality of the subvectors as well as the codebook sizes and their impact on the recognition performance, we considered several permutations of these two parameters for each task. Given the results presented in the previous section, we adopted a LLD-based splitting strategy, and so, each subvector is quantized using the same codevector length and their own codebook.

Figs. 3 and 4 depict the classifier performance for the short-, medium-, and long-term recognition tasks for various codevector lengths (the length of each codevector is $L = log_2 N$, where N is the codebook size) and subvector sizes (k; increasing values of k indicate higher compressions rates). The horizontal lines in each figure indicate the baseline performance for each task. As expected, for increasing codevector lengths (i.e., smaller quantization error) and lower subvector dimensionalities (i.e., lower compression rates) the recognition performance is improved for all tasks, except some cases of the "Pathology" task ($k = 5$ and $k = 175$; discussed below). Taking the "Emotion" task as a representative example (see Fig. 3a), we can observe that for $k = 24$ the UAR varies between (approximately) 62.6 percent ($L = 3$) to 67.0 percent ($L = 12$), a value very close to the baseline (67.6 percent). If we increase the subvector dimensionality (e.g., $k = 48$), the performance varies



(a) Emotion: AEC (b) Intoxication: ALC (c) Pathology: NCSC

Fig. 3. Performance for distributed *short-term* (*emotion*, AEC) and *medium-term* (*intoxication*, ALC; *pathology*, NCSC) paralinguistic tasks. k: subvector size.



(a) Age: Agender (b) Gender: Agender (c) Age+Gender: Agender

w/o VQ × k = 3 ⋆ k = 8 ▫ k = 15 ▪ k = 30 ○ k = 75 • k = 150

Fig. 4. Performance for distributed *long-term* paralinguistic tasks: *age*, *gender*, and *age* + *gender*. k: subvector size.

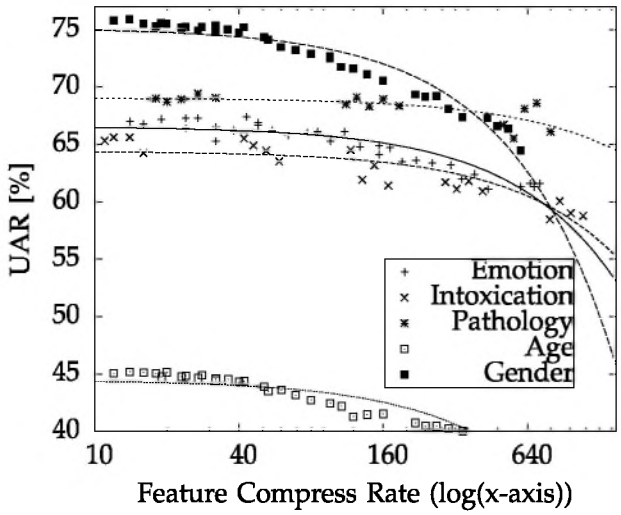


Fig. 5. Relationship between recognition performance (UAR) and *feature compression rate* for the various tasks with manifold permutations of codevector length and subvector dimensionality.

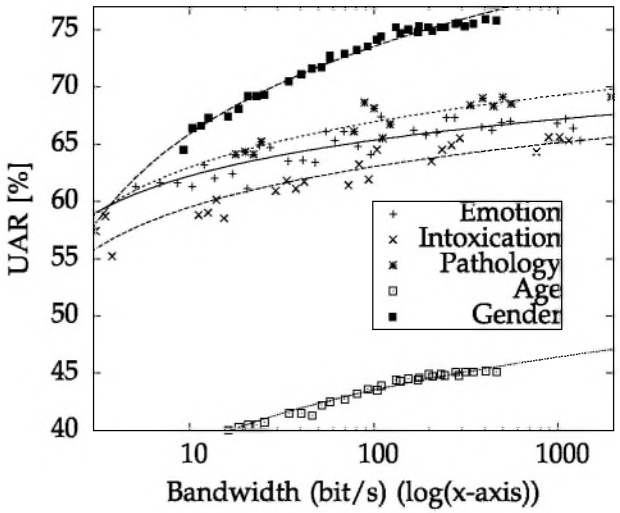


Fig. 6. Relationship between recognition performance (UAR) and *bandwidth requirements* for the various tasks with manifold permutations of codevector length and subvector dimensionality.

between 61.0 percent ($L = 3$) to (approximately) 65.3 percent ($L = 12$), which is further away from the baseline. Naturally, with a higher value of k a smaller bandwidth is required. In the example given, for a codevector of length 12, the bandwidth would be $(384/24) * 12/1.7 \approx 113b/s$ ($k = 24$) and $(384/48) * 12/1.7 \approx 57b/s$ ($k = 48$). Compared to the no compression case the bandwidth reduction would be of 98.4 and 99.2 percent, respectively.

As mentioned above, the Pathology task does not follow the same pattern and shows a more complex relationship between the code vector length and subvector dimensionality. As it can be seen in Fig. 3c for different values of k the performance either decreases ($k = 5$ and $k = 175$) or increases ($k = 35$ and $k = 875$) for increasing code vector lengths. In our view, this phenomenon might be caused by data scarcity. As it can be observed in Table 3, there are only 2,386 instances in total for this task, which is potentially an insufficient number of instances to train a robust SVQ model and/or recognizer. This seems to be corroborated by the results of the "Agender" task, where we have 53,074 instances, and the stability and reliability of the system is much higher (and also the fact that age and gender recognition tasks have a more solid ground truth). Despite this unexpected result, and as it will be shown in the next section, the relationship between recognition performance, feature compression rate, and bandwidth follows a pattern that is similar to that of other tasks (see Figs. 5 and 6). Finally, it is also noticeable that in this task compressing the feature set to a certain degree increased the performance of the model over the baseline—in the case of $k = 5/L = 6$ the UAR is 69.4 percent, and in the case of $k = 35/L = 9$ the UAR is 69.1 percent. This may indicate that, to a certain degree (both values are actually not significantly higher than the baseline), the compression process attributed more weight to relevant features and reduced the impact of less relevant ones.

5.2.1 Performance, Feature Compression Rate, and Bandwidth

Fig. 5 provides a combined overview of the relationship between performance and feature compression rate for the five distributed paralinguistic classification presented in this paper. The distributed gender recognition seems to be the most sensitive to the feature compress rate, as it can be inferred from the slope of the trend line (dash-dot). It can also be observed that, for all tasks except "age", the performance degradation was not significant (one-side z-test, $p > .05$) with feature compression rates smaller than 40, but it is increasingly pronounced for values over 40, and especially over 160. This is an interesting result given that in a multi-task scenario where the best permutation of the subvector dimension and codebook sizes for a given task may be unknown and varied, guaranteeing a compression rate below 40 warrants a good performance for all tasks.

Given that a crucial aspect of a distributed recognition system is the trade-off between recognition accuracy and bandwidth limitations, in Fig. 6 we show the relationship between recognition performance and required transmission bandwidth for all five tasks (the transmission bandwidth is calculated by Eq. (4)). This figure can be used to obtain an estimation of the recognition task accuracy for a particular transmission bandwidth, and vice versa. As expected, the performance decreases for lower transmission bandwidth rates, and is particularly degraded for rates below 100 bit/s. For instance, considering the "Gender" classification task, if a transmission bandwidth of 10 bit/s is available the recognition accuracy would be of about 65.0 percent. If a better performance is required, for instance 75.0 percent, then a transmission bandwidth of more than 100 bit/s would be necessary.

6 Conclusion and Outlook

In this paper, we introduced a general distributed architecture for paralinguistic speech signal processing. We have described the various components of the proposed system, and focused on creating the conditions for large-scale data collection, security of data transmission for protecting personal information, continuous classification improvement and transmission bandwidth optimization. In order to warrant data privacy, we have used statistical features extracted from low-level-descriptors due to their irreversibility.

Furthermore, such strategy also promoted efficiency in terms of performance and bandwidth requirements. The required bandwidth was further optimized by means of features compression. To that end, we have focused on split vector quantization due to its efficiency, security, and the fact that it is the official compression method recommended by ETSI for distributed speech recognition. Finally, we conducted various experiments to investigate the feasibility and efficiency of the system on large-scale paralinguistic tasks, including short-term states, medium-term phenomena, and long-term traits.

We started by showing that there is a strong influence of feature attributes on the performance of the compression algorithm. Compared to a random clustering strategy, grouping the feature attributes under same LLDs reduced the information loss when implementing compression of the feature set using SVQ. We have also shown that subvector and codebook size have a critical impact on the system's performance—the classification performance degrades for almost all tasks when either large subvector or small codebook size (or both) is used. Overall, our results demonstrated that when the feature compression rate is less than 40, the classification performance is similar to that with no compression.

Overall, our results are very informative and encouraging for future exploitation of the system proposed in this paper. Nonetheless, this work is only a first step towards the creation of large-scale distributed paralinguistic information recognition systems for the application in real life contexts, and several issues still need to be addressed. A central issue is the optimization of the various modules. In this paper we focused on demonstrating the feasibility of the whole system, but there is plenty of room for improvements in the various modules. For instance, we have used a common feature compression technique (SVQ), but given the demonstrated importance of the compression stage it would be very beneficial to explore other state-of-the-art feature compression techniques such as principle component analysis [35], linear discriminant analysis [36], histogram-based quantization [14], and sparse representations [59]. Further, while we used preselected features sets for each task, it would be worth exploring the use of feature selection as it could improve compression rates and reduce the required bandwidth while maintaining or improving the recognition tasks performance. Furthermore, given that the dimensionality of statistical features vectors used in this paper is always the same per turn, the transmission bandwidth will vary as a function of turn duration, which may lead to bandwidth bursts for consecutive short turns. A possible way of overcoming such a problem is to consider different methods for dealing with long and short turns so as to avoid its negative impact on the client-server communication. Yet, another possible solution would be to evaluate the contribution of the features used for each classification task, and vary the dimensionality and codebook size for the attributes with different levels of importance (in this paper we considered that all features are equally important to the classification tasks).

In addition to optimization issues there are various important challenges particular to paralinguistic recognition systems that should be addressed in the future. A central one is dealing with multiple paralinguistic tasks simultaneously, and particularly task selection and multi-task learning. In relation to the first, if the tasks are not selected manually on the client side (as considered in this paper) methods such as computational auditory scene analysis could be used to analyze the characteristics of the acoustic environment and infer the paralinguistic task(s) of interest. Concerning the second, and given that it has been continuously demonstrated that paralinguistic tasks benefit from contextual knowledge (for example, gender, social background, and other information can improve emotion recognition performance [60], [61]), it would be relevant to exploit the use of mutual information in multi-task learning scenarios to improve the performance for a particular task. Furthermore, given the overlaps between the feature sets used for different tasks, it is plausible to pursue a common set of features that can be applied to all tasks. In this case the distributed framework can be simplified since modules related to task selection are not necessary anymore. To this end, deep neural networks or sparse coding could be used to extract high-level feature representations which may be shared by the various paralinguistic tasks.

Another aspect that should be at the very center of future developments is the enhancement of the system robustness, in particular regarding recording devices disparity, environmental noise and reverberation, voice variation across users, security protection, and the impact of packet loss during data transmission. This is crucial for the implementation and use of the system proposed in this paper. Furthermore, other technical aspects such as energy efficiency on the client side must be investigated so as to warrant the applicability of the system. Energy optimization algorithms such as optimal time-resource allocation [62] provide ways of defining costs models of local computing and transmission which can be used to find an optimal balance between performance, complexity, computational power and energy consumption.

Despite the many issues still to be addressed in this area, we have shown very promising results and demonstrated that we are not far from the creation of robust distributed multi-task paralinguistic recognition systems that can be applied to a myriad of everyday life scenarios, such as, remote medicine treatments, remote conferences or negotiation, remote education, or even advanced driver assistance systems. Also, and very importantly, as we mentioned in the introduction to this paper, this type of systems may also be crucial to the future of computational paralinguistics by providing the essential speech signals for the development of robust recognition systems.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Research Council in the European Community's 7th Framework Program under grant agreements No. 338164 (Starting Grant iHEARu), and 230331 (Advanced Grant PROPEREMO). It was further partially supported by research grants from the China Scholarship Council (CSC) awarded to Zixing Zhang and Jun Deng. The authors would also thank to Jürgen Geiger for his feedback on an early version of this paper.

REFERENCES

[1] B. Schuller, "The computational paralinguistics challenge," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 97–101, Jul. 2012.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language—State-of-the-art and the challenge," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 4–39, 2013.

[3] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 151–167, 2013.

[4] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2010, pp. 2794–2797.

[5] A. Arimond, "A distributed system for pattern recognition and machine learning," M.Eng. thesis, TU Kaiserslautern & DFKI, Kaiserslautern, Germany, 2010.

[6] Z. Tan and I. Varga, "Network, distributed and embedded speech recognition: An overview," in *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, Z. Tan and B. Lindberg, Eds. Berlin, Germany: Springer, 2008, pp. 1–23.

[7] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 8505–8509.

[8] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE Workshop Autom. Speech Recog. Understanding*, 2011, pp. 523–528.

[9] Z. Zhang, and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, p. 4.

[10] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy, "Active learning from crowds," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 1161–1168.

[11] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 184–198, Apr.–Jun. 2012.

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states—A review on intoxication, sleepiness and the first challenge," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 346–374, 2014.

[13] A. Gomez, A. Peinado, V. Sanchez, and A. Rubio, "Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1228–1238, Dec. 2006.

[14] C. Wan and L. Lee, "Histogram-based quantization for robust and/or distributed speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 859–873, May 2008.

[15] R. Flynn, and E. Jones, "Robust distributed speech recognition using speech enhancement," *IEEE Trans. Consum. Electron.*, vol. 54, no. 3, pp. 1267–1273, Aug. 2008.

[16] Wikipedia. (2014, Feb. 10) Siri—Wikipedia, the free encyclopedia [Online]. Available: http://en.wikipedia.org/wiki/Siri

[17] A. Yang, R. Jafari, S. Sastry, and R. Bajcsy, "Distributed recognition of human actions using wearable motion sensor networks," *J. Ambient Intell. Smart Environ.*, vol. 1, no. 2, pp. 103–115, 2009.

[18] G. Nagy, "Interactive, mobile, distributed pattern recognition," in *Proc. 13th Int. Conf. Image Anal. Process.*, 2005, pp. 37–49.

[19] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Extended Advanced Front-End Feature Extraction Algorithm; Compression Algorithms; Back-End Speech Reconstruction Algorithm*, ETSI ES 202 212 V1.1.1, ETSI standard, 2003.

[20] W. Han, Z. Zhang, J. Deng, M. Wöllmer, F. Weninger, and B. Schuller, "Towards distributed recognition of emotion in speech," in *Proc. 5th Int. Symp. Commun. Control Signal Process.*, 2012, pp. 1–4.

[21] A. Batliner, and B. Schuller, "More than fifty years of speech processing—The rise of computational paralinguistics and ethical demands," in *Proc. ETHICOMP*, Paris, France, 2014.

[22] T. S. Rappaport, *Wireless Communications: Principles and Practice*, vol. 2, Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.

[23] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 3, pp. 1299–1302.

[24] B. Milner and X. Shao, "Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end," *Speech Commun.*, vol. 48, no. 6, pp. 697–715, 2006.

[25] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin, Germany: Springer, 2009.

[26] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 312–315.

[27] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 3201–3204.

[28] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 1–4.

[29] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[30] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 153–158, Feb. 1997.

[31] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[32] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manage.*, vol. 45, no. 3, pp. 315–328, 2009.

[33] S. Planet and I. Iriondo, "Comparative study on feature selection and fusion schemes for emotion recognition from speech," *Int. J. Interactive Multimedia Artif. Intell.*, vol. 1, no. 6, pp. 44–51, 2012.

[34] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit—Searching for the most important feature types signalling emotion-related user states in speech," *Comput. Speech Lang.*, vol. 25, no. 1, pp. 4–28, 2011.

[35] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer, 1986.

[36] Z. Fan, Y. Xu, and D. Zhang, "Local linear discriminant analysis framework using sample neighbors," *IEEE Trans. Neural Netw.*, vol. 22, no. 7, pp. 1119–1132, Jul. 2011.

[37] A. Majumdar, "Distributed face recognition," in *Face Recognition: Methods, Applications and Technology*, A. Quaglia, and C. M. Epifano, Eds. Hauppauge, NY, USA: Nova, 2012.

[38] M. McLaren, and D. van Leeuwen, "Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 755–766, Mar. 2012.

[39] S. S. Kajarekar, B. Yegnanarayana, and H. Hermansky, "A study of two dimensional linear discriminants for ASR," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 137–140.

[40] A. Gersho, and R. M. Gray, *Vector Quantization and Signal Compression*, vol. 159, Boston, MA, USA: Kluwer, 1992.

[41] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1588, Nov. 1985.

[42] T. Lookabaugh, E. Riskin, P. Chou, and R. Gray, "Variable rate vector quantization for speech, image, and video compression," *IEEE Trans. Commun.*, vol. 41, no. 1, pp. 186–199, Jan. 1993.

[43] C. Garcia and G. Tziritas, "Face detection using quantized skin color regions merging and wavelet packet analysis," *IEEE Trans. Multimedia*, vol. 1, no. 3, pp. 264–277, Sep. 1999.

[44] G. McLean, "Vector quantization for texture classification," *IEEE Trans. Syst., Man Cybern.*, vol. 23, no. 3, pp. 637–649, May/Jun. 1993.

[45] J. Arrowood and M. Clements, "Extended cluster information vector quantization (ECI-VQ) for robust classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 889–892.

[46] C. Wan and L. Lee, "Joint uncertainty decoding (JUD) with histogram-based quantization (HQ) for robust and/or distributed speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, pp. 125–128.

[47] A. Vasuki and P. Vanathi, "A review of vector quantization techniques," *IEEE Potentials*, vol. 25, no. 4, pp. 39–47, Jul./Aug. 2006.

[48] V. Digalakis, L. G. Neumeyer, and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 1, pp. 82–90, Jan. 1999.

[49] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Berlin, Germany: Springer, 2005.
[50] A. J. Ferreira, "Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2001, pp. 51–54.
[51] P. A. Naylor and N. D. Gaubitch, *Speech Dereverberation*. Berlin, Germany: Springer, 2010.
[52] A. Batliner, S. Steidl, D. Seppi, and B. Schuller, "Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach," *Adv. Human-Comput. Interaction*, vol. 2010, pp. 3:1–3:15, 2010.
[53] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 483–487.
[54] F. Schiel, C. Heinrich, and S. Barfsser, "Alcohol language corpus: The first public corpus of alcoholized German speech," *Lang. Res. Eval.*, vol. 46, no. 3, pp. 503–521, 2012.
[55] L. Van Der Molen, M. van Rossum, A. Ackerstaff, L. Smeele, C. Rasch, and F. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: Clinical outcome measures and patients' views," *BMC Ear Nose Throat Disorders*, vol. 9, no. 10, 2009.
[56] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proc. 7th Int. Conf. Lang. Resources Eval.*, 2010, pp. 1562–1565.
[57] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[58] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.
[59] K. Huang, and S. Aviyente, "Sparse representation for signal classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 609–616.
[60] B. Schuller, M. Wöllmer, F. Eyben, G. Rigoll, and D. Arsic, "Semantic speech tagging: Towards combined analysis of speaker traits," in *Proc. AES 42nd Int. Conf.*, 2011, pp. 89–97.
[61] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. Eur. Signal Process. Conf.*, 2004, pp. 341–344.
[62] G. Thatte, M. Li, S. Lee, B. Emken, M. Annavaram, S. Narayanan, D. Spruijt-Metz, and U. Mitra, "Optimal time-resource allocation for energy-efficient physical activity detection," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1843–1857, Apr. 2011.

Zixing Zhang received the master's degree in telecommunications from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010. He is currently working toward the PhD degree as a researcher in the Machine Intelligence and Signal Processing Group of Institute for Human-Machine Communication at Technische Universität München (TUM), Munich, Germany. He has published more than 20 papers in peer-reviewed journals and conference proceedings. His research interests include machine learning algorithms for intelligent speech analysis.

Eduardo Coutinho received the diploma in electrical engineering and computer sciences from the University of Porto, Portugal, in 2003, and the doctoral degree in affective and computer sciences from the University of Plymouth, United Kingdom, in 2008. He is currently a postdoctoral fellow at the Technische Universität München, an affiliate researcher at the Swiss Center for Affective Sciences (CISA), and an honorary research fellow at the School of Music from the University of Liverpool. His research interests include the link between low-level acoustics and the expression and perception of emotion in music and speech.

Jun Deng received the bachelor's degree in electronic and information engineering from Harbin Engineering University and the master's degree in information and communication engineering from the Harbin Institute of Technology, Heilongjiang, China, in 2009 and 2011, respectively. He is currently working the PhD degree in the Machine Intelligence and Signal Processing Group, Institute for Human-Machine Communication, Technische Universität München (TUM). His research interests include machine learning and speech emotion recognition.

Björn Schuller (M'05) received the diploma in 1999, the doctoral degree for his study on automatic speech and emotion recognition in 2006, and the habilitation in 2012, all in electrical engineering and information technology from TUM. He is a tenured head of the MISP Group at TUM, a senior lecturer at the Department of Computing, Imperial College London in the United Kingdom, CEO of audEERING UG (limited), a visiting professor of HIT in Harbin, China, an associate of CISA in Geneva, Switzerland and Joanneum Research in Graz, Austria. Previously, he was a full professor heading the Institute for Sensor Systems at the University of Passau, Germany, and with the CNRS-LIMSI's Spoken Language Processing Group in Orsay, France. He is best known for his works in advancing Machine Intelligence for Affective Computing. He is the president of the Association for the Advancement of Affective Computing (AAAC), and an elected member of the IEEE Speech and Language Processing Technical Committee (SLTC), IEEE, ACM, and ISCA and coauthored five books and more than 400 publications in the field leading to more than 6,000 citations—his current h-index equals 39. Community service includes his former cofounding Steering Committee membership and current associate editorship for the *IEEE Transactions on Affective Computing* and further associate editorship including the *IEEE Transactions on Neural Networks and Learning Systems*, the *IEEE Transactions on Cybernetics*, the *IEEE Signal Processing Letters*, and Computer Speech and Language. He is a member of the IEEE.