Neurocomputing 456 (2021) 49–60

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification

Ascensión Gallardo-Antolín [a,*], Juan M. Montero [b]

[a] Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés, Madrid, Spain
[b] Speech Technology Group, ETSIT, Universidad Politécnica de Madrid, Avda. de la Complutense, 30, 28040 Madrid, Spain

ARTICLE INFO

Article history:
Received 15 February 2021
Revised 10 April 2021
Accepted 21 May 2021
Available online 25 May 2021
Communicated by Zidong Wang

Keywords:
Speech intelligibility
LSTM
Attention
Acoustic spectrogram
Modulation spectrogram
Fusion

ABSTRACT

Speech intelligibility can be affected by multiple factors, such as noisy environments, channel distortions or physiological issues. In this work, we deal with the problem of automatic prediction of the speech intelligibility level in this latter case. Starting from our previous work, a non-intrusive system based on LSTM networks with attention mechanism designed for this task, we present two main contributions. In the first one, it is proposed the use of per-frame modulation spectrograms as input features, instead of compact representations derived from them that discard important temporal information. In the second one, two different strategies for the combination of per-frame acoustic log-mel and modulation spectrograms into the LSTM framework are explored: at decision level or late fusion and at utterance level or Weighted-Pooling (WP) fusion. The proposed models are evaluated with the UA-Speech database that contains dysarthric speech with different degrees of severity. On the one hand, results show that attentional LSTM networks are able to adequately modeling the modulation spectrograms sequences producing similar classification rates as in the case of log-mel spectrograms. On the other hand, both combination strategies, late and WP fusion, outperform the single-feature systems, suggesting that per-frame log-mel and modulation spectrograms carry complementary information for the task of speech intelligibility prediction, than can be effectively exploited by the LSTM-based architectures, being the system with the WP fusion strategy and Attention-Pooling the one that achieves best results.

1. Introduction

Speech intelligibility refers to the comprehensibility of speech and plays a crucial role in any process where oral communication is involved. Speech, and therefore its understandability level, can be demeaned due to many factors, such as background noise, reverberation or channel distortions. Besides, certain physiological issues can give place to impairments in any of the components of the human speech production system, producing the so-called pathological or disordered voices.

In this paper, we address the problem of automatic classification of the speech intelligibility level in this latter case, and, in particular, for dysarthric voices. Dysarthria [1] is a speech disorder produced by the motor malfunctioning of the speech organs and characterized by the presence of poor phoneme articulation, imprecise transitions between adjacent phonemes, hypernasality, vocal roughness, perturbations in the elocution rate, volume and pitch, broken speech, etc. It has its origin in neurological injuries due to brain tumors, thrombotic/embolic strokes or degenerative illnesses such as Parkinson's Disease (PD). Dysarthria can seriously hamper the communication for patients, which can even lead to psychological problems.

In recent years, the problem of automatic prediction of disordered speech comprehensibility has attracted the attention of numerous researchers. These studies can be categorized into two main groups [2]: *intrusive* or *non-blind* methods, and *non-intrusive* or *blind* approaches. Intrusive techniques are based on the comparison of the utterance to be evaluated to a reference model that represents non-pathological (intelligible) speech. This model is built from healthy data by using techniques such as Gaussian Mixture Models [3] or iVectors [4]. Other works assume that the behaviour of an Speech-To-Text system trained with healthy speech when recognizing disordered utterances can be indicative of the speaker's intelligibility level [5]. In this sense, features derived from the recognizer output, such as the word error rate, have been proposed for this task. The main drawback of this kind of techniques is that large amounts of non-pathological data are required, although some works have recently tackled this issue [2].

* Corresponding author.
E-mail addresses: gallardo@ing.uc3m.es (A. Gallardo-Antolín), juanmanuel.montero@upm.es (J.M. Montero).

https://doi.org/10.1016/j.neucom.2021.05.065
0925-2312/

In contrast to these techniques, non-intrusive methods do not rely on reference speech signals. They usually treat the automatic estimation of speech intelligibility as a regression or classification problem where the whole utterance is assigned to a single label. This kind of systems typically consist of a front-end where a set of acoustic features are computed and a back-end where the classification itself is performed. In this work, we have followed this approach.

Most of the non-intrusive systems described in the literature rely on traditional machine learning methods, such as Linear Discriminant Analysis (LDA) [6–8], Support Vector Machines (SVM) [9,10] or Random Forests [11]. As these techniques do not allow the adequate modeling of temporal signals as speech, it is necessary to compute an ad hoc *utterance-level representation* at the front-end, that somehow summarizes the information contained in the corresponding temporal sequences. For example, as per-frame acoustic log-mel spectrograms[1] reflect the short-term artifacts of pathological speech, several compact features derived from them are commonly used for intelligibility prediction, as the average of mel-frequency delta-energy coefficients [12] or the average of Mel Frequency Cepstrum Coefficients (MFCC) [10]. In the same way, per-frame modulation spectrograms that convey information about long-term temporal dynamics perturbations of disordered speech are not directly used as input features. Instead, it is utilized some utterance-level characteristics derived from them, such as the frequency and amplitude of the modulation spectrum peak [6], the Low-to–High Modulation Ratio (LHMR) [7,8] or the average energy of the modulation spectrogram [10]. Obviously, these summarized parameterizations imply an important loss of information with respect to the corresponding temporal representations, degrading the performance of the whole system.

Recently, the use of *Long Short-Term Memory* (LSTM) networks [13,14], belonging to the area of Deep Learning (DL) methods, is spreading among the audio and speech research community as they are more suitable for the modeling of temporal sequences. In fact, LSTM-based systems have achieved substantial improvements in several audio and speech-related tasks, such as Acoustic Event Detection (AED) [15], Acoustic Scene Classification (ASC) [16], Speech Emotion Recognition (SER) [17,18] or Cognitive Load (CL) classification from speech [19,20]. In many of these works, LSTMs are combined with Weighted Pooling (WP) schemes in order to obtain an utterance-level representation that is subsequently processed by the succeeding dense layers that compose the classifier itself. Nowadays, one of the most successful WP methods is the so-called *attention pooling* [21,17], a mechanism that tries to determine the structure of the temporal sequences by learning the relevance of each frame to the task under consideration. These attentionl LSTM models have been successfully proposed for the aforementioned speech and audio tasks. A detailed review about this topic can be found in [22].

In our previous paper [10], we showed that is feasible to model per-frame acoustic log-mel spectrograms with attentional LSTM networks for the task of predicting the speech intelligibility level (low, medium or high) of a dysarthric person and we achieved significant improvements with respect to a SVM-based system with different types of ad hoc compact acoustic features.

One of the weakness of our previous research is that it was focused on short-term perturbations of pathological speech, whereas long-term disturbances were not taken into consideration. For overcoming this problem, in this paper we extend the work into two directions. Firstly, we propose a similar attentional LSTM architecture for modeling the long-term dynamics of disordered speech. For this purpose, we consider the adoption of the per-frame modulation spectrogram as input features to the network, what, to the best of our knowledge, has not been previously reported for this task. Secondly, as it is reasonable to hypothesize that log-mel and modulation spectrograms contain complementary information for the estimation of a subject's speech intelligibility, we explore two different alternatives for the fusion of both kind of features and show that the combination improves the performance of the individual systems. Again, to the best of our knowledge, the combination of this kind of per-frame features into a LSTM-based architecture for speech intelligibility prediction has not been previously explored in the literature.

The rest of the document is organized as follows: Section 2 describes the motivation behind the use of acoustic log-mel and modulation spectrograms as features for the automatic estimation of the intelligibility level; Section 3 contains the general description of the proposed attention LSTM-based intelligibility classification system, including the single-feature architecture that uses either log-mel or modulation spectrograms as inputs, and the architecture combining both kind of features; the database, experiments and results are described in Section 4; and Section 5 finishes the document with some conclusions and future work.

2. Acoustic log-mel and modulation spectrograms as indicators of speech intelligibility

Several phenomena, such as vocal harshness, nasal voice, adventitious sounds, unclear transitions between adjacent phonemes, excessive phoneme duration, variable speech rate, disfluencies, presence of odd pauses, and other prosodic disturbances (pitch breaks, monotonicity, etc.), are usually present to some extent in dysarthric speech, affecting negatively its intelligibility level [23,7].

Some of these artifacts can be observed by analyzing the speech recordings at short-time scales. In fact, the short-time *acoustic spectrogram*, which shows the variation with time of the speech spectrum computed at short *acoustic frames*, provides useful cues about the intelligibility degree of an utterance [24]. The acoustic spectrogram of a speech signal $x(n)$ is computed by performing a short-time analysis using the following Equation [24],

$$X_a(n_a, k'_a) = \sum_{l=-\infty}^{\infty} x(l) w_a(n_a - l) e^{-j\frac{2\pi k'_a}{N_a} l} \tag{1}$$

where the subscript a denotes the acoustic domain, n_a is the frame index and it is related to the discrete-time index n through the expression $n_a = \frac{n}{L_a}$, being L_a the frame period (in samples); k'_a is the index of the acoustic frequency bin; and $w_a(n)$is the analysis window whose length is N_a (in samples).

Prior studies have concluded that the energies of certain frequency subbands suffer variations due to the presence of distorted phonemes and that those subbands placed in the limits of the F1-F2 space are more affected to vowels' pronunciation problems [25,12]. In addition, the amount of energy above $3500Hz - -4000Hz$ in dysarthric speech is usually less than in non-pathological speech [26,12].

Log-mel spectrograms are derived from the corresponding acoustic spectrograms that are first mapped to the mel-frequency spacing [27] by using an auditory filter bank composed of mel-scaled filters, and later converted to a logarithmic scale. Log-mel spectrograms are denoted by $S_a(n_a, k_a)$, where k_a is the index of the mel-scale filter and their computation process is represented in Fig. 1. One of their advantages is that they are bio-inspired spectro-temporal representations of speech, as the mel scale is a frequency warping that tries to mimic the non-equal sensitivity of the human hearing at different frequencies. As previously men-

[1] Throughout this paper, we use indistinctly the terms "acoustic log-mel spectrogram", "log-mel spectrogram" and "log-mels".

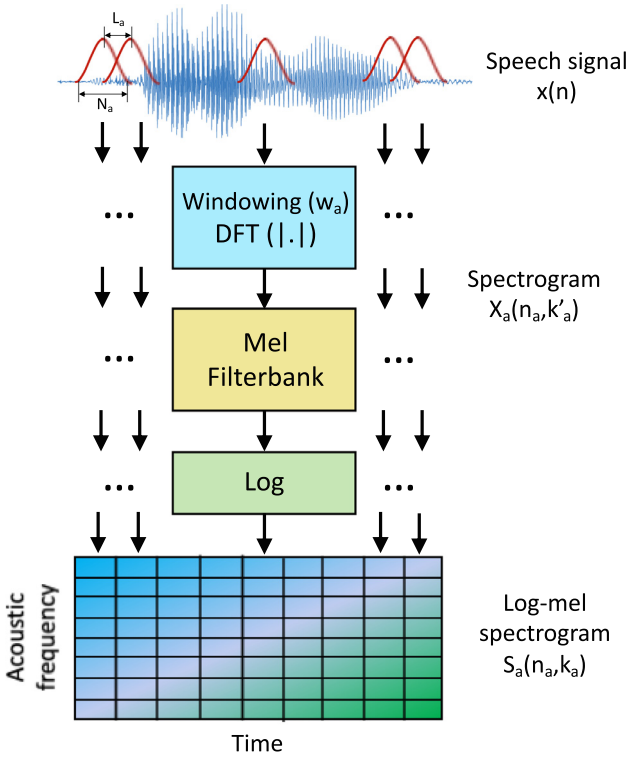

Fig. 1. Block diagram of the log-mel spectrogram computation process.

tioned, when using traditional machine learning algorithms, log-mel spectrograms are not directly used as features. Instead, a compact representation, such the average of mel-frequency delta-energy coefficients [12], the standard deviation of the first derivative of the zeroth order MFCC [7,8] or the average of MFCC [10] are utilized, what could bring a relevant loss of information.

For overcoming this issue and following our previous work [10] where it was proven that per-frame log-mel spectrograms are suitable for the task of intelligibility prediction in combination with LSTM-based classifiers, in this paper, we have chosen to use this kind of features for coping with the short-term characteristics of pathological speech. In particular, for the log-mel spectrograms computation, we have used a Hamming window with length $N_a = 20\ ms$, a frame period of $L_a = 10\ ms$ and a mel-scale filterbank composed of 32 filters.

Dysarthric speech characteristics related to rhythmic disturbances are better reflected in representations that capture long-term speech temporal dynamics information. In particular, features derived from the *modulation spectrogram* of speech [28,29], which captures the speed of fluctuation of long-term speech temporal envelopes, are good candidates for this purpose. This is because it has been shown that slow temporal envelope modulations contain information about several perturbations that can be present in pathological speech, such as, non–habitual intensity and speed variations, imprecise coarticulations or interruptions and disfluencies, and, therefore they can be used as indicators of speech intelligibility [30,24,7].

From the analysis of the speech modulation spectrogram properties, previous studies have shown that in healthy voice most of the energy is located in the range of modulation frequencies from $2-20\ Hz$, reaching maximum values at around $4\ Hz$, which is the modulation frequency corresponding to the average syllabic rate [28]. However, in dysarthric voice the peak of the modulation energy suffers a shift to lower values. In addition, as the intelligibility degree decreases, the bulk of modulation energy tends to concentrate in lower modulation frequencies [6,7].

In order to illustrate these previous findings, we have analyzed the modulation spectrograms of the utterances contained in the database used in our experiments (see SubSection 4.1). The modulation spectrograms are obtained following the method proposed in [30,31] that is summarized in Fig. 2. First, the speech signal $x(n)$ is decomposed into acoustic frequency bands by means of a filterbank composed of 23 gammatone critical-band filters, resulting in a set of 23 filtered signals $x_k(n), k = 1, \ldots, 23$. This filterbank mimics the cochlear processing that takes place in the human auditory system and its first and last filter are centered at 125 Hz and 8000 Hz, respectively. Then, the temporal envelopes corresponding to each frequency band, $e_k(n)$, are computed by using the Hilbert transform $H\{.\}$, as follows,

$$e_k(n) = \sqrt{x_k(n)^2 + H\{x_k(n)^2\}} \tag{2}$$

Then, the modulation spectrum for a given acoustic frequency k is computed as the Discrete Fourier Transform of the corresponding temporal envelope $e_k(n)$, as indicated in the following equation,

$$X_m(n_m, k, k'_m) = \sum_{l=-\infty}^{\infty} e_k(l) w_m(n_m - l) e^{-j\frac{2\pi k'_m}{N_m} l} \tag{3}$$

where the subscript m denotes the modulation domain, n_m is the frame index and it is related to the discrete-time index n through the expression $n_m = \frac{n}{L_m}$, being L_m the frame period (in samples); k'_m is the index of the modulation frequency bin; and $w_m(n)$is the analysis window whose length is N_m (in samples). In this work, $w_m(n)$ is a Hamming window with length $N_m = 256ms$ and the frame shift is $L_m = 64ms$. The modulation frequencies k'_m are grouped according to a modulation filterbank composed of 8 s-order bandpass filters with a quality factor $Q = 2$, whose center frequencies are located in the range $2-64\ Hz$, yielding the modulation spectrogram that is denoted as $S_m(n_m, k, k_m)$, where k_m is the index of the modulation filter. Note that for each *modulation frame* n_m, it is obtained a 2D representation that contains the modulation energies, whose dimensions are the number of acoustic frequencies multiplied by the number of modulation frequencies (in this case, 23×8).

Fig. 3 shows two examples of the modulation spectrum of a particular modulation frame belonging to an utterance with high (a) and low intelligibility (b). In both cases, the horizontal and vertical axes represent, respectively, the modulation and acoustic frequencies. It can be observed that for the low intelligibility speaker, the modulation energies are concentrated in low modulation frequencies, being the maximum values located below $4\ Hz$ (see Fig. 3(b)), whereas in the case of the high intelligibility speaker, the energy peak is situated around $4Hz$ and the modulation energy spreads over higher modulation frequencies (see Fig. 3(a)).

Fig. 4(a) shows the average relative increment/decrement of energy per modulation band for low and medium intelligibility utterances with respect to the energy of the corresponding band of the high intelligibility recordings. These values were computed from one of the folds of the dataset used in our experimentation (see SubSection 4.1). Energies per modulation band were computed by averaging, for each modulation band, the energies across the different acoustic frequencies and normalizing them by the total modulation energy of the utterance. As this figure shows, in the case of low intelligibility, low-frequency modulation energy is slightly increased with respect to high intelligibility speech, whereas in the range of approximately $3Hz-10Hz$, the opposite occurs. This fact corroborates the observations drawn in the aforementioned studies [6,7]. For medium intelligibility, same trends are observed although the increment/decrement values are smaller than in the previous case. For all intelligibility levels, the absolute energy located in very high-frequency regions is small, so its

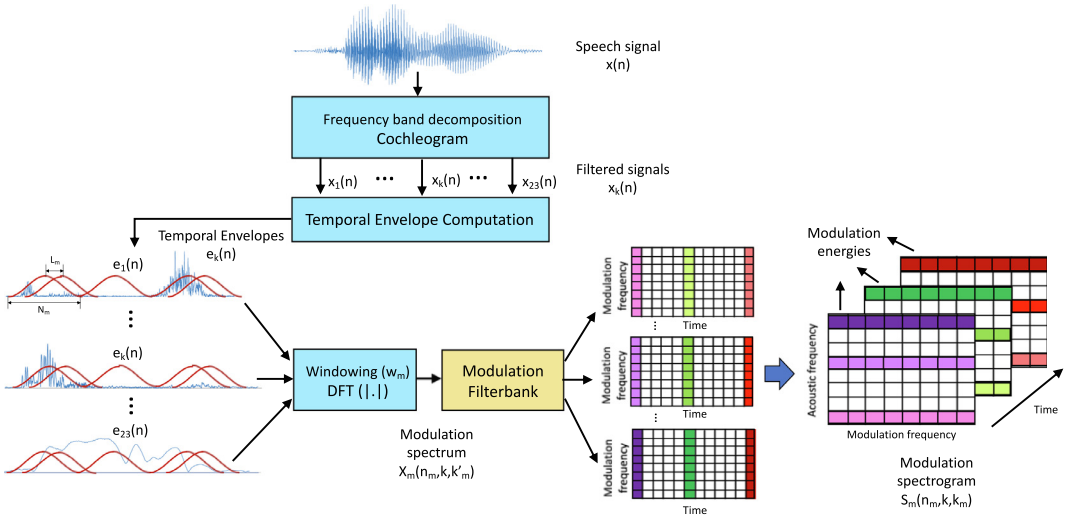

Fig. 2. Block diagram of the modulation spectrogram computation process.

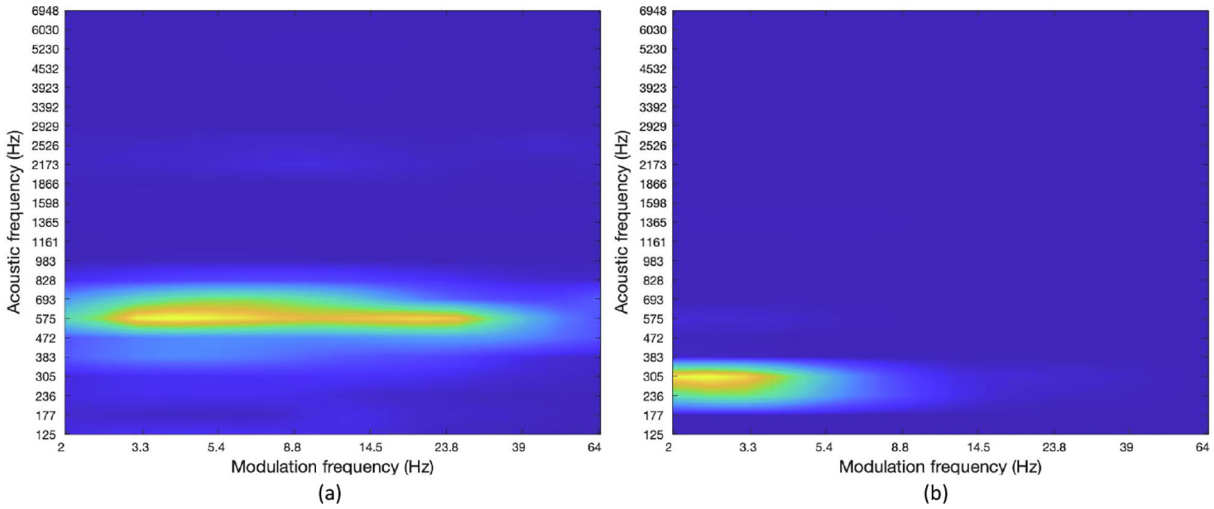

Fig. 3. Modulation spectrum of a certain modulation frame belonging to a speech recording with (a) high intelligibility and (b) low intelligibility. Both frames correspond to the central part of the first vowel of the word "jowls".

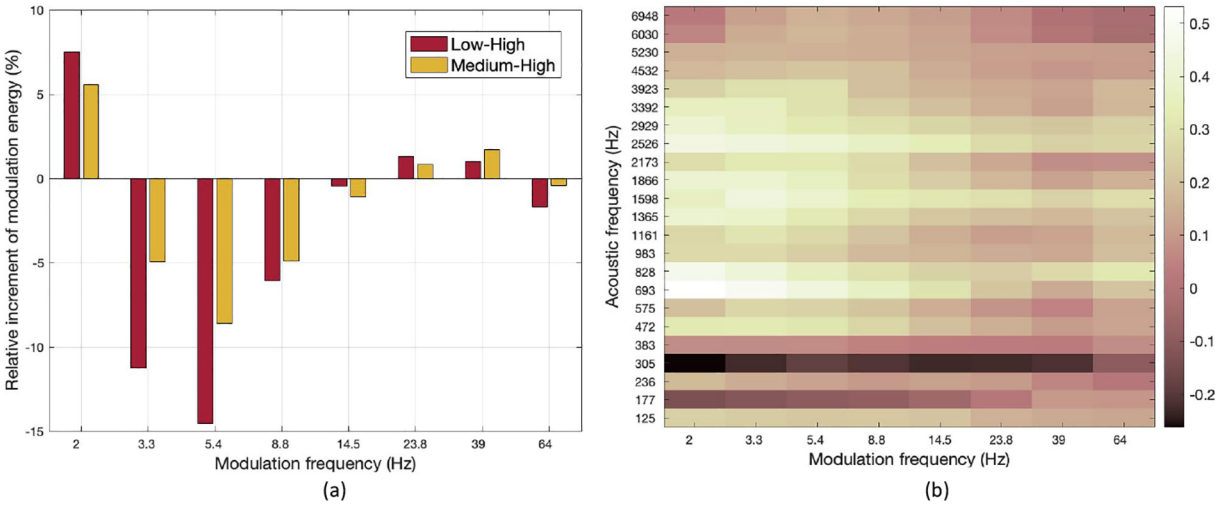

Fig. 4. Relationship between modulation energy and speech intelligibility level: (a) Average relative increment/decrement of energy per modulation band for low and medium intelligibility utterances with respect to the energy of the corresponding band of the high intelligibility recordings; (b) Correlation values between modulation energy and intelligibility level.

relative increments or decrements do not seem to provide a significant information for our task.

Fig. 4(b) represents the correlation values between the per-frame modulation energy and the intelligibility level. These values have been calculated from the same fold of the dataset than in the previous case. On the one hand, it can be seen that the region that is directly correlated with the intelligibility degree is roughly the one with a modulation frequency below 10 *Hz* and an acoustic frequency from 450 *Hz* to 3500 *Hz*. On the other hand, the intelligibility level seems to be slightly inversely correlated with energies at acoustic frequencies around 300 *Hz* independently of the modulation frequency. Although, in general, the individual correlations are not very high, it seems apparent that the modulation spectrogram at a whole contains useful cues for the inference of the intelligibility, what justifies the use of automatic learning techniques able of exploiting this information.

In previous research, it has been proposed several features derived from the modulation spectrogram for speech intelligibility prediction in combination with traditional machine learning classifiers, as LDA [6–8] or SVMs [9,10]. The Low-to-High Modulation energy Ratio (LHMR), which is the quotient of modulation energy at modulation frequencies below 4 *Hz* to modulation frequencies above 4 *Hz* [6,7], frequency and amplitude of the modulation spectrum peak, energy in the region of $3--6$ *Hz* [6] or the average of modulation energies across the frames [10] are good examples. However, as these features can be seen as summarizations of the modulation spectrogram content, their use likely implies a loss of temporal information.

Regarding this issue, our conjecture is that, as in the case of log-mel spectrograms, DL-based models are able to exploit all the information conveyed in the per-frame modulation spectrograms, allowing their effective use in an automatic intelligibility classification system. In particular, due to the temporal nature of the per-frame modulation spectrograms, we propose the use of LSTMs for this purpose, as explained in Section 3.

Besides, recalling that dysarthric speech characteristics are related to short and long-term phenomena, we argue that a combination of features extracted at different time scales is required in order to better determine the intelligibility level of an utterance. For this reason, we propose the fusion of log-mel and modulation spectrograms in the framework of a LSTM-based system as described in SubSection 3.2.

3. LSTM-based speech intelligibility classification system

Long Short-Term Memory networks are a special type of Recurrent Neural Networks that are able to learn long-term dependencies due to their capacity to store past information in their memory blocks [13,14], in such a way that their outputs depend on past and present inputs. For that, LSTMs are very suitable for modeling temporal sequences, such as log-mel and modulation spectrograms that are the feature inputs considered in this work.

LSTMs perform a sequence-to-sequence learning where an input sequence of length $T, x = \{x_1, \ldots, x_T\}$ is converted to an output sequence $y = \{y_1, \ldots, y_T\}$ of the same length. As in the speech intelligibility classification problem, a single label (intelligibility level) must be assigned to the whole input sequence, a Weighted Pooling (WP) stage is connected to the LSTM layer, whose purpose is to aggregate the information contained in y and produce an utterance-level representation z that, in turn, is the input to the classifier itself [17,32]. WP is commonly implemented as a simple weighted aggregation operation as follows,

$$z = \sum_{t=1}^{T} \alpha_t y_t \tag{4}$$

where $y = \{y_1, y_2, \ldots, y_T\}$ is the LSTM output sequence, $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ is the weight vector and z is the final utterance-level representation. The way weights are computed determines different types of WP approaches. In this work, we have considered the following ones.

Basic LSTM. In this case, it is assumed that the last frame of the LSTM output, y_T, is the most representative one, as information from the whole sequence has been used to some extent for its computation [33]. Therefore, only this last frame goes into the succeeding layer, discarding the remaining ones, i.e., $z = y_T$. In other words, all the weights are zero except the one corresponding to the last LSTM frame that is equal to one (i. e., $\alpha_t = 0, t = 1 \ldots, T-1$ and $\alpha_T = 1$).

Mean-Pooling. In the Mean-Pooling approach, the utterance-level representation is calculated as the average of the output LSTM frames along the whole sequence, as indicated in the following expression,

$$z = \frac{1}{T}\sum_{t=1}^{T} y_t \tag{5}$$

In this case, it is assumed that all the LSTM frames are equally important. Therefore, all the weights are equal ($\alpha_t = \frac{1}{T}, \forall t$) and, as a consequence, all the elements of y contribute evenly to z.

Attention-Pooling. The rationale behind this approach is that not all the LSTM frames reflect the understandability level with the same intensity. For this reason, frames containing more cues about the intelligibility degree should be more emphasized than the remaining ones. This way, the utterance-level representation z is computed as the weighted arithmetic mean of the output LSTM frames, where larger weights should be set to the more relevant frames to the task, whereas smaller weights should be assigned to frames not conveying useful information.

For the computation of the weights, we have adopted the method proposed by [18] for speech emotion recognition, that is especially suitable for scenarios where the amount of training data is limited, as is our case. In this approach, the unnormalized attention weights, $\bar{\alpha} = \{\bar{\alpha_1}, \bar{\alpha_2}, \ldots, \bar{\alpha_T}\}$, are calculated through the following Eq. (6).

$$\bar{\alpha_t} = u^{tr} y_t, \tag{6}$$

where u and y are, respectively, the attention parameter vector and the LSTM output, and the superscript tr denotes a transpose operation. Note that the inner product between u and y_t measures the relevance of each t-th frame. In order to obtain a set of normalized weights whose sum across all the frames of the sequence is equal to one, a softmax transformation is applied to these quantities, according to the following expression,

$$\alpha_t = \frac{\exp(\bar{\alpha_t})}{\sum_{t=1}^{t=T} \exp(\bar{\alpha_t})} \tag{7}$$

The attention parameters and the LSTM outputs are obtained in the training process of the network.

In the following Subsections, we describe the architectures proposed for the single-feature LSTM-based system, where using only log-mel spectrograms or modulations spectrograms as input, and the combined LSTM-based systems, where both, log-mel and modulations spectrograms are jointly utilized as input features.

3.1. Single-feature LSTM-based architecture

Fig. 5(a) shows the block diagram of the single-feature LSTM system proposed for speech intelligibility classification, which is based on our previous work [10]. As can be seen, the general architecture of the system is the same, no matter the input consists of

Table 1
Values of the configuration parameters for the LSTM-based systems.

Parameter	Description	Value	
		Log-mel spectrogram	Modulation spectrogram
T	No. of frames of the input signal	Variable – Range: [121, 5691]	Variable – Range: [19, 890]
n_F	No. of features in the log-mel/ modulation spectrogram per frame	32	184
L	No. of frames of the LSTM input/ output sequences	700	110
n_{D1}	No. of neurons in the first dense layer	32	100
n_L	No. of LSTM units	64	64
n_{D2}	No. of neurons in the second dense layer	25	25
n_C	No. of classes (intelligibility levels)	3	3

log-mel or modulation spectrograms. However, the specific values of the variables in the model are different depending of the kind of input features considered, as is detailed in Section 4 and Table 1.

The dimension of the input features x_{input} are $T \times n_F$, being T the number of acoustic or modulation frames of the speech signal and n_F the number of components of the feature vectors. In the case of log-mel spectrograms n_F matches the number of mel filters, whereas in the case of modulation spectrograms n_F is equal to the number of acoustic frequencies multiplied by the number of modulation frequencies considered in the modulation spectra computation.

As both, log-mel and modulation spectrograms do not have the same length for all the speech recordings and, however, the length of the LSTM input sequences is set to a fixed value L, a preprocessing of the sequences is required in order to overcome this issue. In particular, longer sequences than L are cut (this only occurs in a few cases as it is shown in SubSection 4.2), whereas shorter sequences are padded with masked values. A masking layer allows that these dummy values not to be used in further computations. Note that the value of L is different for each type of feature. This is because log-mel and modulation spectrograms are extracted at different frame periods as indicated in SubSection 4.2.

Next, a dense layer of n_{D1} neurons is connected to the masking layer. Its purpose is to perform some kind of feature extraction for improving the input to the following LSTM layer. This consists of n_L cells, being the output of each cell a sequence of length L. The information of these LSTM sequences is compressed by using a Weighted Pooling stage. As aforementioned, three strategies for WP have been considered: basic LSTM, Mean-Pooling and Attention-Pooling.

The WP module is connected to a second dense layer of n_{D2} neurons with dropout in order to avoid overfitting. In the end, its output goes into a final dense layer with n_C nodes that is activated by a softmax function for performing the multiclass classification. The value of n_C matches the number of possible intelligibility levels to be predicted (low, medium and high).

3.2. Fusion strategies for the combination of log-mel and modulation spectrograms

We have explored two fusion strategies for the combination of log-mel and modulation spectrograms, namely, *late fusion* and *WP fusion*, as depicted in Fig. 6. In the first case, the combination is produced at decision level by means of a dense layer with n_C neurons and sigmoid activation whose inputs are directly the final outputs of the individual systems. In the second case, the fusion is performed at utterance level, i.e., the outputs of the Weighted Pooling layers of the individual systems are combined by using a dense layer of $2 \times n_{D2}$ nodes with dropout, followed by a final fully-connected layer of n_C neurons and sigmoid activation.

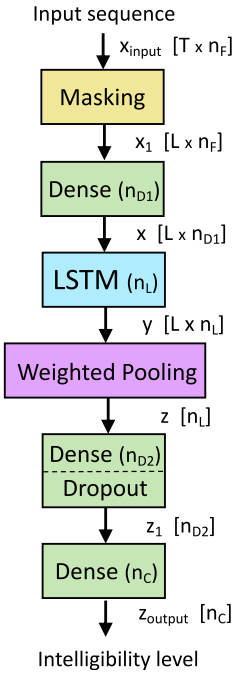

Fig. 5. LSTM-based architecture for speech intelligibility classification when using either log-mel spectrograms or modulation spectrograms as input features. The dimensions of each variable are indicated in brackets.

4. Experiments

4.1. Database

For our experiments, we have used the UA-Speech database [34] that consists of utterances (digits, computer commands, simple short words, complex long words and the radio alphabet) pronounced by 15 persons (11 men and 4 women) suffering from dysarthria with different degrees. All the speech files were recorded at 16*KHz* with an array of 7 microphones, although only signals corresponding to the sixth microphone have been utilized. The dataset also contains speech from healthy control subjects. However, these audio recordings have been discarded, as in this work we have followed the non-intrusive approach for the design of the speech intelligibility classification system. In summary, the total number of available files for the experiments is 11,435.

The database was manually annotated by medical staff in terms of the intelligibility score, that can be defined as the average percentage of understood words by the specialists after carrying out a series of subjective tests. As intelligibility scores range from 0 (completely unintelligible) to 100 (perfectly intelligible), these original labels were mapped to the three categories considered in this work. This way, the low, medium and high intelligibility classes correspond to, respectively, scores from 0 to 33, from 34 to 66 and from 67 to 100.

Regarding the experimental protocol, we have used a subject-wise 5-fold cross validation. Specifically, the database was split into five disjoint balanced groups, in such a way that all recordings from the same subject were included in the same subset. In each fold, one group was kept for testing, another different group was utilized for validation, whereas the remainder ones were used for

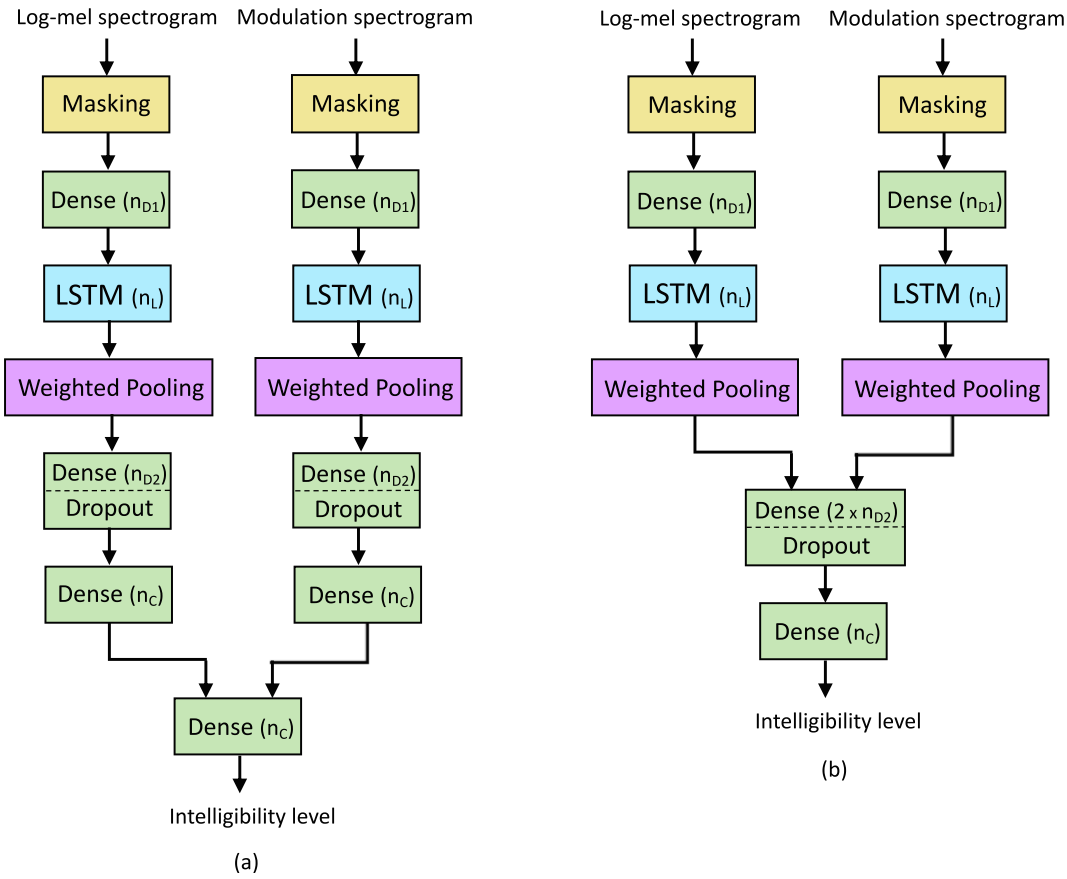

Fig. 6. Fusion strategies for the combination of log-mel and modulation spectrograms. n_{D1}, n_L, n_{D2} and n_C stand for the number of neurons in the first dense layer, the number of LSTM units, the number of neurons in the second dense layer of the individual systems and the number of classes (intelligibility levels), respectively. (a) Late fusion: combination at decision level; (b) WP fusion: combination at utterance level.

training. The experiments were repeated five times alternating the training, validation and test sets, averaging the results afterwards. We adopt this speaker-independent configuration in order to prevent the system to learn the speaker's identity instead of his/her intelligibility level.

4.2. Feature extraction

Log-mel spectrograms were computed using a Hamming window of duration $N_a = 20\ ms$ with a frame shift of $L_a = 10\ ms$ and a filterbank composed of 32 triangular filters distributed according to the mel scale. For this purpose, the Python's package LibROSA [35] was utilized. Mean and standard deviation normalization were applied at utterance-level obtaining a set of normalized log-mel spectrogram sequences x_{input} with $T \times n_F$ dimensions, where T is the number of acoustic frames of each utterance (note that each acoustic frame corresponds to $10ms$) and n_F is the number of mel filters, i.e., $n_F = 32$.

Modulation spectrograms were computed over windows of duration $N_m = 256\ ms$ with a frame shift of $L_m = 64\ ms$ using 23 critical band frequencies and 8 modulation filters according to the recommendations in [7,8,31], and by using the Matlab software SRMR [30]. For each modulation frame, a 2D modulation spectrum was calculated and subsequently flattened, producing a vector of 23 x 8 = 184 dimensions. Again, mean and standard deviation normalization were applied at utterance-level yielding a set of normalized modulation spectrogram sequences x_{input} with $T \times n_F$ dimensions, being, in this case, T the number of modulation frames (note that each modulation frame corresponds to $64ms$) and n_F the dimension of each vectorized modulation spectrum, i.e., $n_F = 184$.

For the efficient computation of LSTM networks, it is required that all the input sequences have the same dimensions, and therefore, a fixed length was established for all audios. In particular, the maximum length was set to $7s$ because more than 95% of the audio signals were shorter than this quantity (see [10] for details). This duration corresponds to $L = 700$ and $L = 110$ for, respectively, log-mel and modulation spectrograms. Feature sequences longer than L were cut, and, otherwise, they were padded with masked values that are ignored in further computations by means of the use of the appropriate masking layers in the LSTM-based architectures.

4.3. LSTM-based classifiers

All the individual LSTM-based systems (Basic, Mean-Pooling and Attention-Pooling) for the two kind of features (log-mel and modulation spectrograms), as well as the corresponding fused systems, were implemented with the Python's packages Tensorflow [36] and Keras [37]. The specific values of the configuration parameters of the single and combined architectures depicted, respectively, in Figs. 5 and 6 are detailed in Table 1. In all cases, the LSTM models were trained using stochastic gradient descent and the Adam optimization algorithm with an initial learning rate of 0.0002, a batch size of 32 and a maximum number of 50 epochs. In order to avoid overfitting in the training process, a 33% dropout was set in certain layers as indicated in Figs. 5 and 6.

In the Attention-Pooling systems, the attention parameter vector u had a dimension of $n_L = 64$ and its components were initialized to $1/L$, i.e., to $1/700$ and $1/110$ for log-mel and modulation spectrograms, respectively.

4.4. Reference systems

For comparison purposes, several reference systems based on the traditional machine learning technique SVM were implemented using the MATLAB Statistics and Machine Learning toolbox.

Firstly, two single-feature SVM-based systems were developed considering two different types of compact parameterizations: the average of the MFCC and the average of the modulation spectrogram. The first parameterization, MFCC, is a very popular feature extraction procedure in audio and speech related tasks (see, for example, [38,31]), and for this reason, it was tried for the task under consideration in our previous work [10]. MFCCs are extracted on a frame-by-frame basis by applying the Discrete Cosine Transform on the log-mel spectrogram of the speech signal (see SubSection 4.2) and retaining the first 13 coefficients. These coefficients are augmented by adding their first derivatives. Finally, the average of these parameters over all the acoustic frames are computed, yielding a final vector composed of 26 components per utterance. The second parameterization consists of the average across all the modulation frames of the energies of the modulation spectrogram, which is obtained following the procedure mentioned in SubSection 4.2. In this case, each utterance is represented by a 184-dimension vector.

Secondly, two more SVM-based systems were implemented for assessing the combination of the two aforementioned sets of acoustic characteristics. In the first case, systems were fused at decision level (*late fusion*). In particular, the scores produced by each of the individual SVM systems were transformed to probabilities by applying a softmax operation, and then multiplied between them for obtaining the final score. In the second case, the fusion was done at feature level by means of the concatenation of the average of MFCC and the average of the modulation spectrogram into a single feature vector (*early fusion*).

In all systems, a Radial Basis Function (RBF) kernel was used and the optimal hyperparameters of the model were obtained by means of a Bayesian optimizer with a 5-fold cross validation strategy.

4.5. Results

The developed systems were assessed in terms of the *accuracy or classification rate per audio recording*, that is defined as the percentage of correctly classified files with respect to the total number of tested files. In all cases, each experiment was run 20 times and therefore, results reported in the tables contained in this Section are given as the average accuracy across the 20 subexperiments together to the corresponding standard deviation.

*Results of the individual systems.*Table 2 contains the accuracies achieved by the single-feature LSTM-based systems with the two different parameterizations under consideration: log-mel and modulation spectrograms, and the three Weighted Pooling schemes studied in this work: Basic, Mean-Pooling and Attention-Pooling. For comparison purposes, classification rates obtained by the reference SVM-based systems with the average of MFCC and with the average of the modulation spectrograms as input features are also reported.

In order to analyze the statistical significance of the results, Fig. 7 depicts the recognition rates achieved by all the systems evaluated in this work, together to the corresponding 95% confidence intervals.

Table 2
Average classification rates [%] and the corresponding standard deviations achieved by the SVM-based reference system with either average of MFCC or the average energy of the modulation spectrogram as input features, and the single-feature LSTM-based classifiers with either log-mel spectrograms or modulation spectrograms as input features.

System	Features	Accuracy [%]
SVM	Average MFCC	41.68 % ± 0.85 %
SVM	Average modulation spec.	45.81 % ± 0.71 %
Basic LSTM - No Pooling	Log-mel spectrogram	61.35 % ± 0.28 %
LSTM Mean-Pooling	Log-mel spectrogram	65.77 % ± 0.47 %
LSTM Attention-Pooling	Log-mel spectrogram	**68.93 % ± 0.35 %**
Basic LSTM – No Pooling	Modulation spectrogram	61.38 % ± 0.26 %
LSTM Mean-Pooling	Modulation spectrogram	64.39 % ± 0.36 %
LSTM Attention-Pooling	Modulation spectrogram	**67.33**% ± **0.54**%

Firstly, it can be observed that the performance of the reference SVM-based system is poor regardless of the input features, and significantly worse than any of the LSTM-based systems. This result suggests the importance of correctly modeling the temporal behavior of the acoustic features.

Secondly, the classification rates obtained by LSTM when using log-mel spectrograms as input features are analyzed. Note that these results are not directly comparable to those reported in our previous paper [10] because here we adopted a 5-fold cross validation strategy as mentioned in SubSection 4.1, whereas in [10] a fixed dataset partition (50% for training, 15% for validation and 35% for test) was used. As can be observed, the simplest WP strategy, (*Basic LSTM – No Pooling*), which corresponds with the case where only the last LSTM frame is considered for classification, produces the worst result. The most plausible explanation is that with this approach useful information contained in the remaining LSTM frames is ignored. In fact, results improve when the Mean-Pooling strategy (*LSTM Mean-Pooling*) is applied, as, in this case, all the LSTM frames contribute with the same weight to the utterance-level representation. Nevertheless, further improvements can be achieved when incorporating the attention mechanism into the LSTM framework (*LSTM Attention-Pooling*), since it allows the system to learn what are the most relevant parts of the utterances with regard to speech intelligibility estimation, diminishing the contribution of the remaining ones. In particular, *LSTM Attention-Pooling* obtains a relative error reduction of 19.61% with respect to *Basic LSTM – No Pooling* and of 9.23% with respect to *LSTM Mean-Pooling*.

Thirdly, as for the use of modulation spectrograms with LSTM, results show that this kind of features contain some important cues about the comprehensibility of an utterance than can be effectively exploited by means of LSTM-based classifiers. Regarding the performance of the different WP approaches, again, the *LSTM Attention-Pooling* system clearly outperforms the other two WP strategies, achieving relative error reductions of 15.41% and 8.26% with respect to *Basic LSTM – No Pooling* and *LSTM Mean-Pooling*, respectively. This behavior suggests that it is important to emphasize the contribution of the more significant frames for the task independently of the type of features used.

Finally, regarding the comparison between the two kind of parameterizations in the LSTM framework, results with log-mel spectrograms are very similar to those achieved by modulation spectrograms for the *Basic LSTM - No Pooling* system. However, for the *LSTM Mean-Pooling* and *LSTM Attention-Pooling* approaches, log-mels perform better than modulation spectrograms, although the differences are not very noticeable. In fact, they are not statistically significant as is observed in Fig. 7. This result suggests that modulation spectrograms can be used as an alternative parameterization to log-mels. Moreover, in next paragraphs, it will be shown

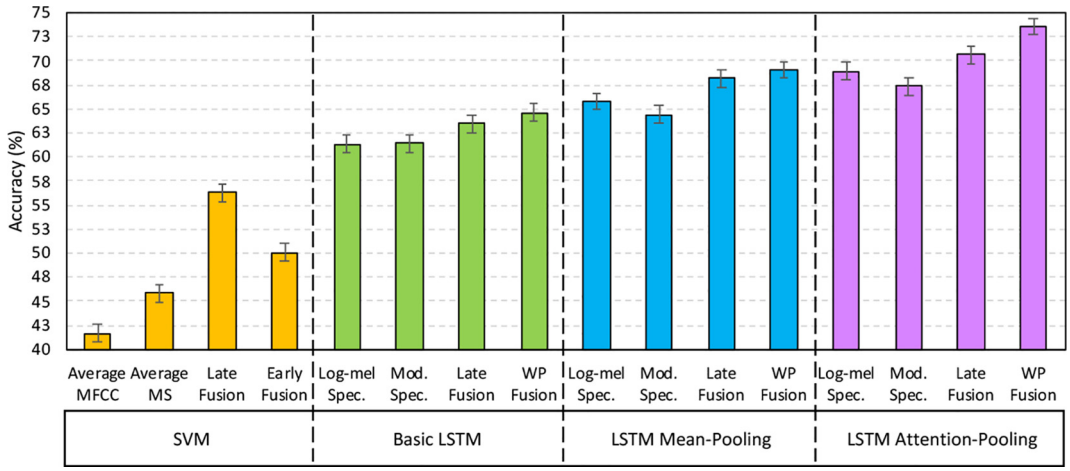

Fig. 7. Average classification rates [%] and the corresponding 95% confidence intervals for both, SVM-based and LSTM-based systems.

that both kind of features can be utilized in a complementary manner.

As an example of the behavior of the attention mechanism, Fig. 8(a) depicts the waveform (top) of an utterance with high intelligibility and the corresponding Mean-Pooling and Attention-Pooling weights when using log-mels as features (middle) and when using modulation spectrograms (bottom). Fig. 8(b) displays the same information for a low intelligibility case. As a general remark, for all cases, in contrast to the uniformity of Mean-Pooling weights, Attention-Pooling ones show a significant variation with time, depending on the relevance to the task that this mechanism assigns to each temporal frame. Attention weights for both kind of features exhibit similar trends although the weight curves corresponding to modulation spectrograms are smoother due to the larger analysis windows used for their computation in comparison to log-mels. From the analysis of these graphs, it can be observed that, although larger weights are assigned to high energy speech segments, low energy frames corresponding to pauses or speech artifacts also contribute to the final utterance representation as their weights are greater than zero. For example, the segment from 0.5 to 0.8*s* in the low intelligibility utterance corresponds to a hesitation and the corresponding attention weights present high values. This fact suggests the importance of frames related to disfluencies or rhythmic disturbances for the determination of the intelligibility level in both, log-mel and modulation domains.

Results of the fused systems. Table 3 contains the results attained by the systems where the two type of features, log-mel and modulation spectrograms, are combined. In particular, the corresponding classification rates for the two fusion strategies proposed in this work: fusion at decision level (*Late Fusion*) and at utterance level (*WP Fusion*), and the three WP methods under consideration: Basic, Mean-Pooling and Attention-Pooling, are reported. It also contains the accuracies achieved by the combined SVM-based systems for both, late and early fusion. Together to these classification rates, Fig. 7 shows the corresponding 95% confidence intervals.

In the case of SVM, both fusion strategies outperform the corresponding single-feature systems, being *late fusion* the method that produces best results. However, these accuracies are significantly worse than those attained by any of the individual and fused LSTM-based systems.

Analyzing the behaviour of the fused LSTM-based systems, it can be observed that both combination approaches improve the performance of the individual systems for the three WP schemes. As in the case of the single-feature systems, *LSTM Attention-Pooling* is the best option followed by *LSTM Mean-Pooling*. These results corroborate our hypotheses that log-mel and modulation spectrograms carry complementary information about the level of intelligibility of an utterance and that the attention mechanism is an useful tool for this task, no matter the features used as input to the LSTM-based systems.

The comparison between the two fusion strategies in LSTM-based systems shows that the combination at utterance level is more beneficial than the mixture of the outputs of the respective single-feature systems. Specifically, the best system (*WP Fusion + Attention-Pooling*) achieves a relative error reduction of 39.41% with respect to *SVM + Late Fusion*, 14.84% with respect to *Log-mel spectrograms + Attention-Pooling*, of 19.01% with respect to *Modulation spectrograms + Attention-Pooling* and of 9.97% with respect to the second best system (*Late Fusion + Attention-Pooling*). In addition, the differences in performance between *WP Fusion + Attention-Pooling* and all the remaining systems are statistically significant, as can be observed in Fig. 7.

In order to perform a more detailed analysis about the complementarity of log-mel and modulation spectrograms, we have obtained the confusion matrices produced by the single-feature systems *Log-mel spectrograms + Attention-Pooling* and *Modulation spectrograms + Attention-Pooling*, and the fused system *WP Fusion + Attention-Pooling*. These confusion matrices are represented in, respectively, Fig. 9(a)–(c). In the three graphics, the rows correspond to the correct class, the columns to the hypothesized one and the values in them are computed as the average over the test utterances belonging to the fourth fold.

As can be observed, in the case of log-mel spectrograms, the less confusable class (with a classification rate greater) is high intelligibility, whereas there is about 37% of low intelligibility utterances that are misclassified as medium ones. For modulation spectrograms, the class presenting the best performance is low intelligibility. In this case, more than 40% of medium intelligibility items are incorrectly assigned to the low class. This distinctive behaviour supports the hypothesis that both sets of features carry complementary information, so their combination can outperform the individual systems, as is corroborated by the results in Table 3 and Fig. 7, and the confusion matrix of the fused system represented in Fig. 9.

4.6. Computation complexity

In this Subsection, the computational complexity of the LSTM-based systems developed in this work is analyzed. Firstly, for all WP schemes, the sequence learning performed by the LSTMs is the process that has more impact on the overall computational

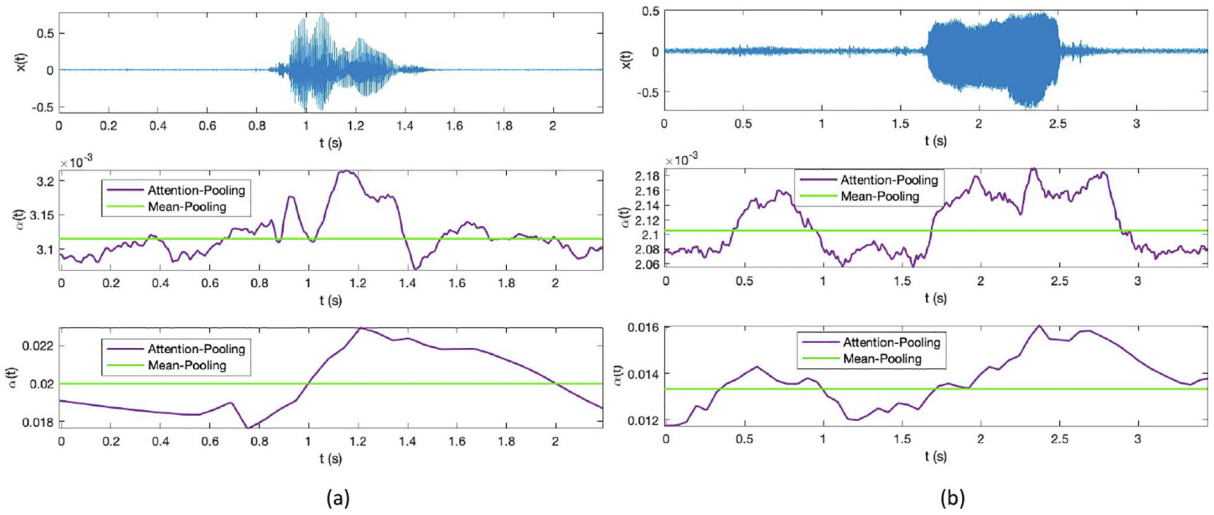

Fig. 8. Mean-Pooling and Attention-Pooling weights for an utterance with (a) high intelligibility and (b) low intelligibility. Top: Waveform. Middle: Weights corresponding to the Mean-Pooling (green line) and Attention-Pooling approaches (violet line) for the LSTM-based system with log-mel spectrograms as input features. Bottom: Weights corresponding to the Mean-Pooling (green line) and Attention-Pooling (violet line) approaches for the LSTM-based system with modulation spectrograms as input features. Both utterances correspond to the word "jowls".

Table 3
Average classification rates [%] and the corresponding standard deviations achieved by the SVM-based reference system with two types of combination strategies: late and early fusion, and the LSTM-based classifiers with two types of combination strategies: late fusion and WP fusion.

System	Type of combination	Accuracy [%]
SVM	Late Fusion	56.33 % ± 0.79 %
SVM	Early Fusion	50.10 % ± 0.75 %
Basic LSTM – No Pooling	Late Fusion	63.47 % ± 0.67 %
LSTM Mean-Pooling	Late Fusion	68.16 % ± 0.61 %
LSTM Attention-Pooling	Late Fusion	**70.61**% ± **0.52** %
Basic LSTM - No Pooling	WP Fusion	64.60 % ± 0.51 %
LSTM Mean-Pooling	WP Fusion	69.00 % ± 0.55 %
LSTM Attention-Pooling	WP Fusion	**73.54 %** ± **0.48 %**

cost. For that, we ignore the effect of the remaining stages when estimating the complexity of the whole systems.

According to [13], the LSTM algorithm is very efficient, exhibiting a time complexity per time step of $O(W)$, where W is the number of trainable parameters (weights). Therefore, the overall complexity of a single LSTM-based system is $O(W \times L)$, where L is the LSTM input sequence length. The complexity of the combined systems can be approximated as the sum of the complexity of the corresponding individual systems.

Table 4 contains the values of W, L and $O(.)$ for the LSTM-based systems with log-mel and modulation spectrograms as input features and the fused system. It also contains the relative complexity with respect to the system that uses log-mel spectrograms as input features. As can be observed, when utilizing modulation spectrograms, the complexity is four times less than for log-mels, and, therefore it is a good alternative in cases where there are computational constraints. Otherwise, the complexity of the combined system is about 1.27 times that of log-mels. This increase in complexity is acceptable, taking into account the benefits achieved in terms of accuracy.

4.7. Limitations of the study

There are two major limitations in this work that could be addressed in future research. The first is the lack of diversity of the training data, in the sense that the number of speakers (15 speakers) and intelligibility scores (15 different scores in the range from 0 to 100) is scarce. In addition, the data is imbalanced with respect to the genre of the speakers (11 men

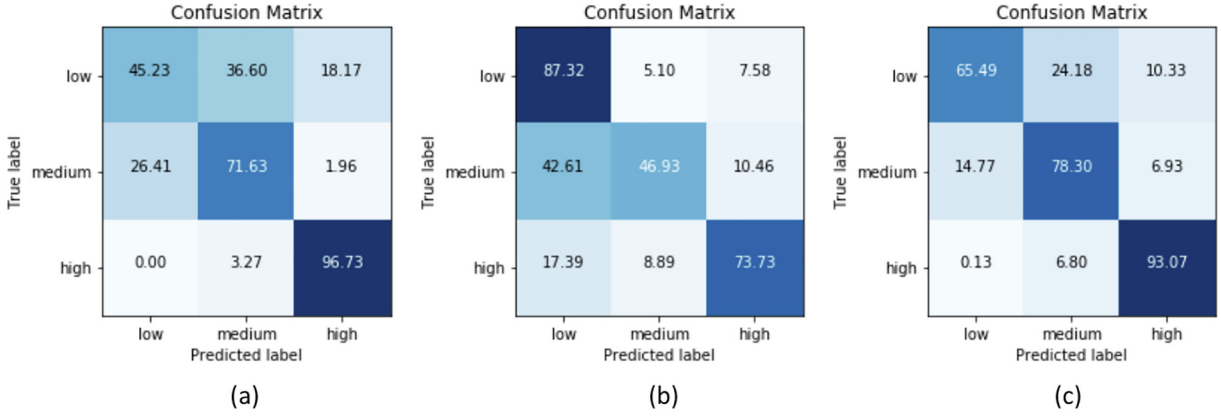

Fig. 9. Confusion matrices [%] for the LSTM-based systems with attention and three different input features: (a) Log-mel spectrograms; (b) Modulation spectrograms; (c) Combined system with WP fusion.

Table 4
Computational complexity of LSTM systems based on log-mel spectrograms, modulation spectrograms and their fusion.

System	W	L	$O(.)$	Relative complexity
Log-mel spectrogram	24832	700	16.58 M	1
Modulation spectrogram	42240	110	4.43 M	≈ 0.27
Combination	67072	–	21.01 M	≈ 1.27

vs. 4 women). Due to the difficulty of collecting data for this task, the exploration of the appropriate data augmentation methods could be an interesting option for overcoming, at least partially, this issue.

The second limitation concerns the accuracy achieved by the best of our systems, that implies that around 25% of the audio files are incorrectly classified. On the one hand, classification rates could be improved by means of the use of the aforementioned data augmentation techniques. On the other hand, it must be taken into account that speech intelligibility classification is a challenging task, not only for machines but also for humans, and depends to some extent on the intrinsic characteristics of the utterance to be listened. In this sense, the audio recordings in the database contain utterances of very different types: digits, radio alphabet, computer commands, simple short words and multisyllable complex words. The characteristics of these broad groups of utterances are rather different regarding, among others, their duration, pronunciation difficulty, and phoneme similarity and confusability. An exhaustive analysis of the classification errors as a function of the aforementioned factors could give insight into the design of the best set of words to be used for intelligibility measurement.

5. Conclusions and future work

In this paper, we have extended our previous work about the development of an automatic non-intrusive system for speech intelligibility level classification based on attention LSTM networks. We present two main contributions. In the first one, we have proposed the use as input features of per-frame modulation spectrograms, instead of compact representations derived from them that discard important temporal information. As modulation spectrograms capture long-term phenomena typically presented in pathological speech, whereas log-mel spectrograms are more related to short-term events, in the second contribution we have explored two different strategies for the combination of both kind of per-frame features into the LSTM-based architecture: at decision level (late fusion) and at utterance level (WP fusion). In both cases, three different weighting schemes in the LSTM architecture have been evaluated: basic LSTM, LSTM with Mean-Pooling and LSTM with Attention-Pooling.

The developed systems have been assessed over the UA-Speech database that contains dysarthric speech with different levels of severity. First, it can be observed that LSTM-based systems significantly outperform the performance of traditional SVM-based systems. Second, results have shown that attention LSTM networks are able to adequately modeling the modulation spectrograms sequences producing similar classification rates as in the case of the LSTM-based system with log-mel spectrograms. In all cases, the weighting scheme based on the attentional mechanism is the one that achieves the best results, so it is clear that learning the contribution of each frame to the task improves the system performance. Third, both combination strategies, late and WP fusion, outperform the single-feature systems, suggesting that log-mel and modulation spectrograms carry complementary information than can be effectively exploited by the LSTM-based architectures. Best results have been achieved by the combined system with WP fusion and Attention-Pooling, obtaining relative error reductions of 39.41 % with respect to SVM + Late Fusion, and of 14.84% and 19.01% with respect to log-mel spectrograms and Attention-Pooling and modulation spectrograms and Attention-Pooling, respectively. All these performance differences are statistically significant.

Regarding the applicability of this study, the assessment of the speech intelligibility level can be useful, among others, as part of diagnosis aid systems for certain diseases or the monitoring of patients following logopedic or other medical therapies. This is, for example, the case of PD where the Unified Parkinson's Disease Rating Scale (UPDRS), which is commonly used to track the progression of this illness, contains two assessment criteria related to the speech communication skills of the patient [39]. The gold standard for judging the comprehensibility level of a patient's speech consists of carrying out several tests where the subject pronounces various words and/or sentences. Then, one or more specialists listen to these utterances and assign the intelligibility scores according to the percentage of words they understood. Nevertheless, multiple benefits can be achieved from the automatization of this task, as is proposed in this work. Firstly, it allows more time to the medical staff to carry out other activities. Secondly, the specialists' criterion to assess the intelligibility is in a way subjective as it relies on their own hearing skills that might be influenced by their familiarization with pathological speech [40]. Finally, it provides an objective and reproducible measure.

For future work, we plan to study the use of weights derived from auditory saliency features [41] in the attention mechanism, following the promising results achieved in [19] for cognitive load estimation from speech. In addition, we plan to extend our research towards the exploration of the more suitable data augmentation techniques for improving the training process and performance of the speech intelligibility level classification system. Finally, the analysis of the classification errors as a function of several factors, such as utterance duration, pronunciation difficulty and phoneme confusability could help to select the most appropriate set of words for intelligibility measurement.

CRediT authorship contribution statement

Ascensión Gallardo-Antolín: Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. **Juan M. Montero:** Conceptualization, Methodology, Software, Investigation, Supervision, Writing - review & editing, Visualization, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work leading to these results has been partly supported by the Spanish Government-MinECo under Projects TEC2017-84395-P and TEC2017-84593-C2-1-R. The authors wish to acknowledge Dr. Mark Hasegawa-Johnson for making the UA-Speech database available.

References

[1] P. Doyle, H. Leeper, A.-L. Kotler, N. Thomas-Stonell, C. O'Neill, M.-C. Dylke, K. Rolls, Dysarthric speech: A comparison of computerized speech recognition and listener intelligibility, Journal of Rehabilitation Research and Development 34 (1997) 309–316.

[2] P. Janbakhshi, I. Kodrasi, H. Bourlard, Spectral subspace analysis for automatic assessment of pathological speech intelligibility, Proc. Interspeech (2019) 3038–3042.

[3] T. Bocklet, K. Riedhammer, E. Nöth, U. Eysholdt, T. Haderlein, Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling, Journal of Voice 26 (2012) 390–397.

[4] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, A. Miguel, Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace, ACM Transactions on Accessible Computing 6 (2015).

[5] A. Zlotnik, J.M. Montero, R. San-Segundo, A. Gallardo-Antolín, Random Forest-Based Prediction of Parkinson's Disease Progression Using Acoustic, ASR and Intelligibility Features, Proc. Interspeech (2015) 503–507.

[6] J.M. Liss, S. LeGendre, A.J. Lotto, Discriminating dysarthria type from envelope modulation spectra, Journal of Speech, Language, and Hearing Research 53 (2010) 1246–1255.

[7] T.H. Falk, W.-Y. Chan, F. Shein, Characterization of atypical vocal source excitation, temporal blackdynamics, and prosody for objective measurement of dysarthric word intelligibility, Speech Communication 54 (2012) blue622–631.

[8] M. Sarria-Paja, T.H. Falk, Automated dysarthria severity classification for improved objective intelligibility assessment of spastic dysarthric speech, Proc. Interspeech (2012) 62–65.

[9] T. Khan, J. Westin, M. Dougherty, Classification of speech intelligibility in parkinson's disease, Biocybernetics and Biomedical Engineering 34 (2014) 35–45.

[10] M. Fernández-Díaz, A. Gallardo-Antolín, An attention long short-term memory based system for automatic classification of speech intelligibility, Engineering Applications of Artificial Intelligence 96 (2020) 103976.

[11] H. Byeon, Developing a model for predicting the speech intelligibility of south korean children with cochlear implantation using a random forest algorithm, International Journal of Advanced Computer Science and Applications 9 (2018).

[12] R. Hummel, W.-Y. Chan, T.H. Falk, Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech, Proc. Interspeech (2011) 3017–3020.

[13] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

[14] F.A. Gers, N.N. Schraudolph, J. Schmidhuber, Learning precise timing with LSTM recurrent networks, Journal of Machine Learning Research 3 (2003) 115–143.

[15] Kao, C.-C., Sun, M., Wang, W., Wang, C., A comparison of pooling methods on lstm models for rare acoustic event classification, in: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)..

[16] Guo, J., Xu, N., Li, L.-J., Alwan, A., Attention based cldnns for short-duration acoustic scene classification, in: Proc. Interspeech 2017..

[17] C.-W. Huang, S.S. Narayanan, Attention assisted discovery of sub-utterance structure in speech emotion recognition, Interspeech (2016) 1387–1391.

[18] Mirsamadi, S., Barsoum, E., Zhang, C., Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231..

[19] A. Gallardo-Antolín, J.M. Montero, A Saliency-Based Attention LSTM Model for Cognitive Load Classification from Speech, Proc. Interspeech (2019) 216–220.

[20] A. Gallardo-Antolín, J.M. Montero, External attention LSTM models for cognitive load classification from speech, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11816 LNA, I (2019) 139–150.

[21] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y. , Attention-based models for speech recognition, in: Proceedings of the 28th International Conference on Neural Information Processing Systems – Volume 1, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, p. 577–585..

[22] N. Zacarias-Morales, P. Pancardo, J.A. Hernández-Nolasco, M. Garcia-Constantino, Attention-inspired artificial neural networks for speech processing: A systematic review, Symmetry 13 (2) (2021) 214.

[23] M.S. De Bodt, M.E. Hernández-Díaz Huici, P.H. Van De Heyning, Intelligibility as a linear combination of dimensions in dysarthric speech, Journal of Communication Disorders 35 (2002) 283–292.

[24] K. Paliwal, B. Schwerin, K. Wójcicki, Role of modulation magnitude and phase spectrum towards speech intelligibility, Speech Communication 53 (2011) 327–339.

[25] H.M. Liu, F.M. Tsao, P.K. Kuhl, The effect of reduced vowel working space on speech intelligibility in mandarin-speaking young adults with cerebral palsy, The Journal of the Acoustical Society of America 117 (2005) 3879–3889.

[26] R. Kent, J. Duffy, A. Slama, J. Kent, A. Clift, Clinicoanatomic studies in dysarthria: Review, critique, and directions for research, Journal of Speech, Language, and Hearing Research 44 (2001) 535–551.

[27] P. Mermelstein, Distance measures for speech recognition, psychological and instrumental, Pattern Recognition and Artificial Intelligence 116 (1976) 374–388.

[28] S. Greenberg, B.E.D. Kingsbury, The modulation spectrogram: in pursuit of an invariant representation of speech, IEEE International Conference on Acoustics, Speech, and Signal Processing 3 (1997) 1647–1650.

[29] J. Vicente-Peña, A. Gallardo-Antolín, C. Peláez, F. Díaz-de-María, Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition, Speech Communication 48 (2006) 1379–1398.

[30] T.H. Falk, C. Zheng, W. Chan, A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech, IEEE Transactions on Audio, Speech, and Language Processing 18 (2010) 1766–1774.

[31] M. Sarria-Paja, T.H. Falk, Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification, Computer Speech & Language 45 (2017) 437–456.

[32] Huang, C.-W., Narayanan, S.S., Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition, in: Proc. of ICME 2017, pp. 583–588..

[33] R. Zazo, A. Lozano-Díez, J. González-Domínguez, D.T. Toledano, J. González-Rodríguez, Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks, PLOS ONE 11 (2016) 1–17.

[34] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T.S. Huang, K. Watkin, S. Frame, Dysarthric speech database for universal access research, in: 9th Annual Conference of the International Speech Communication Association (INTERSPEECH), ISCA, 2008, pp. 1741–1744.

[35] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, et al., LibROSA/LibROSA (2020) 0.7.2.

[36] Abadi, M. et al., Tensorflow: Large-scale machine learning on heterogeneous systems, 2015..

[37] F. Chollet et al. https://keras.io.

[38] A. Gallardo-Antolín, R. San-Segundo, UPM-UC3M system for music and speech segmentation, in: Proc. VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, FALA, 2010, pp. 421–424.

[39] C. Goetz, The unified Parkinson's disease rating scale (UPDRS): status and recommendations, Movement Disorders 18 (2003) 738–750.

[40] S. Landa, L. Pennington, N. Miller, S. Robson, V. Thompson, N. Steen, Automatic assessment of speech intelligibility for individuals with aphasia, International Journal of Speech-Language Pathology 16 (2014) 408–416.

[41] E.M. Kaya, M. Elhilali, Modelling auditory attention, Philosophical Transactions of the Royal Society B 372 (2017) 1–10.

Ascensión Gallardo-Antolín received her Ph. D. in Telecommunication Engineering from the Universidad Politécnica de Madrid (UPM) in 2002. She has been predoctoral researcher at UPM and lecturer at Universidad Autónoma de Madrid. She is currently Associate Professor at Universidad Carlos III de Madrid (Dept. of Signal Theory and Communications). She has been a visiting scientist at the International Computer Science Institute (USA), the German Research Center for Artificial Intelligence (Germany) and the Centre for Speech Technology Research (UK). She has coauthored more than 90 peer-reviewed papers in JCR journals and national and international conferences. She has participated in more than 40 research projects (including some of the Spanish Council on Science and Technology and the UE) and/or contracts with technological companies, mainly related to multimedia processing (speech, audio, video). She has received the Best Ph. D. Thesis Award from the Professional Association of Telecommunication Engineers of Spain (COIT) and the Ph. D. Excellence Award from the Universidad Politécnica of Madrid. Her main research interests include signal processing for multimedia human-machine interaction, automatic speech recognition, audio classification and segmentation, auditory and visual salience models and speech-based health applications.

Juan M. Montero is Telecommunication Engineer (1992) and Ph.D. (2003) by Universidad Politécnica de Madrid (UPM). Currently, he is Associate Professor at the Electronic Engineering Department (UPM). Prof. Montero has been visiting researcher at the International Computer Science Institute (U.C. Berkeley, 2005) and at the Centre for Speech Technology Research (University of Edinburgh, 2013). He has authored 140 peer-reviewed papers (more than 30 in JCR-indexed journals). He has been Principal Investigator in 10 national and international research projects on machine learning techniques for speech, language, health and education.