Interspeech 2018
2-6 September 2018, Hyderabad

Detecting Packet-Loss Concealment Using Formant Features and Decision Tree Learning

Gabriel Mittag[1], *Sebastian Möller*[1,2]

[1]Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
gabriel.mittag@tu-berlin.de, sebastian.moeller@tu-berlin.de

Abstract

One of the main quality impairments in today's packet-based voice services are interruptions caused by transmission errors. Therefore, most codecs comprise concealment algorithms that attempt to reduce the perceived quality degradation of missing speech packets. In case the algorithm fails to properly synthesize the lost speech, interruptions or unnatural sounds are usually perceivable by the user. When measuring the quality of a voice network, there are excellent tools available, which can predict the perceived speech quality. However, they offer only little insight into the technical cause of a quality degradation. A packet-loss detection model could explain the influence of transmission errors on the speech quality and state a packet-loss rate. Thus, making it easier to identify technical problems in the network. In this paper, we examine a new approach for detecting (perceived) packet-loss of transmitted speech by audio analysis. After finding a lost packet, the model classifies in a second stage if the loss was perceivable as a quality degradation. In the model, we use meaningful features that are easy to interpret, and obtained promising results in a simulated environment. Therefore, this detector could also be used to evaluate new packet-loss concealment algorithms and help in optimizing the same.

Index Terms: speech quality, packet-loss concealment, plc

1. Introduction

The quality of transmitted speech is usually measured by auditory tests in which test participants rate speech files on a five-point absolute-category scale [1]. The average over all participants then gives the mean opinion score (MOS) of a condition. Because these auditory tests are costly and time consuming, it is usually preferred to use instrumental quality models. The current recommendation for instrumental quality prediction of super-wideband (SWB) speech by the International Telecommunication Union (ITU-T) is ITU-T Rec P.863 or P.OLQA [2]. It estimates the MOS of a speech signal based on a comparison of the original signal and the degraded signal. However, the MOS only gives a statement about the overall speech quality, without any insights into the problem that caused the quality impairment. In [3] three orthogonal speech quality dimensions were identified: Noisiness, Coloration, and Discontinuity; to which later Loudness was added as a fourth dimension [4]. An instrumental prediction model for these dimensions was presented in [4] and is also currently under study at the ITU-T SG12 in the work item P.AMD [5]. These quality dimensions provide more diagnostic information about the speech impairment, but still there is no direct link to a technical cause of the transmission system, since these quality dimension are purely perceptual based. Therefore, the ITU-T SG12 work item P.TCA [6] aims at identifying problems in a communication network through diagnostics and cause analysis of speech signals. Since frame/packet loss or concealment is one of the main impairments in modern voice networks, it is necessary for such a problem identification model to detect these losses through audio analysis. Concealment algorithms of modern codecs try to synthesize a lost packet by repeating information from the previously received frames [7]. When the algorithm fails to synthesize a signal similar to the missed frame, it often results in robotic voice or artificial and annoying sounds. In end-to-end measurements of voice transmission systems not all network parameters, such as packet-loss rate, are usually available. For this application, a signal based packet-loss detection model could estimate a packet-loss rate and explain the influence of transmission errors on the speech quality. Furthermore, it could be used to train single-ended speech quality models, which are not able to predict packet-loss reliably at the moment [8].

In this paper, we study a novel approach for detecting (perceived) packet-loss concealment by using meaningful features, such as formant information and fundamental frequency distances. We then train a first decision tree classifier that detects frames which are affected by packet-loss, in a second stage another tree classifier is trained to find only frames in which the packet-loss was auditory perceivable. As a results we can calculate an estimate of a perceived packet-loss rate and additionally a rate of frames with packet-loss that were notable in the speech signal, but not perceivable. This detection model could then also be used to evaluate or optimize new concealment algorithms. In the experiments, we focused on the SWB codec EVS [9] because it uses state of the art concealment algorithms. However, as we also want to study if our approach works codec independently, we included AMR-WB [10] conditions in the test database as well. In the following we will first introduce the features that are used in the detection model, then we describe the decision tree classifier and the databases that we generated. After that, we present the results and finally conclude the paper.

2. Features

The goal of this detection model is to use meaningful features that can easily be interpreted. Using features that are directly linked to properties caused by packet-loss, should make the detector more robust to other distortions, such as noise and loudness variation. Therefore, we want to avoid using features such as MFCCs or simple distance measurements between the original and the degraded power spectral densities. Most of the packet-loss concealment methods used by CELP codecs are based on extrapolating LP-parameters from previously received frames to conceal the missing frame. If multiple consecutive frames are lost, the gain of the concealed frames is attenuated towards comfort noise level.

Thus, one main indicator of a concealed missed packet is a

10.21437/Interspeech.2018-1098

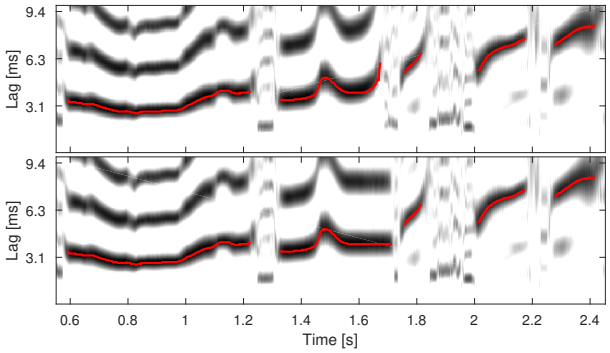

Figure 1: *Autocorrelograms with extracted fundamental frequency (in lags) marked as red line.* TOP*: Original autocorrelogram* $\mathbf{R}_{xx}$. BOTTOM*: Degraded autocorrelogram* $\mathbf{R}_{yy}$.

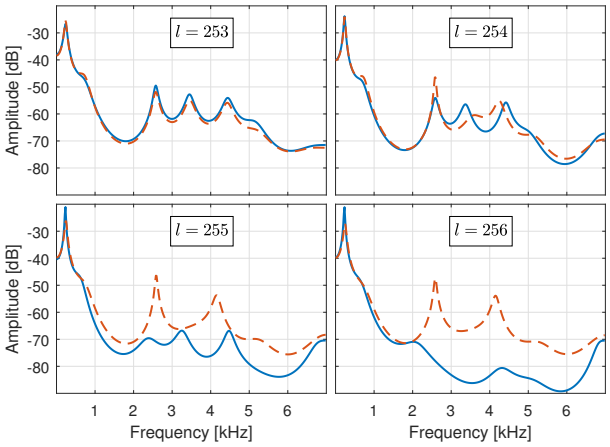

Figure 2: *Spectral envelope of four consecutive speech frames.* SOLID, BLUE*: Original signal.* DASHED, RED*: Degraded signal*

constant fundamental frequency in the degraded signal y, while the fundamental frequency in the original signal x is changing. This effect is shown in Figure 1, where at around 1.7 seconds it can be seen how the fundamental frequency (red line) stays constant in the bottom graph of the degraded signal (measured in lags), whereas the lag of the fundamental frequency increases towards the end of the vowel in the original signal in the top graph. As a consequence, we used the distance in fundamental frequency as a first input feature of the detection model. Although generally a change in the pitch of a speech signal is not necessarily linked to a perceived quality degradation, it almost always seems to be the case for transmitted speech through voice networks, as the pitch change is accompanied by other distortions in the frequency domain.

In some cases the fundamental frequency may stay constant in the degraded signal, but equally so in the original signal. However, the perceived quality of a speech signal will still be clearly compromised by artifacts if the formants of the original signal are severely altered. Figure 2 presents the spectral envelope of four consecutive speech frames. In the first frame, shown in the top left graph, the degraded spectral envelope (dashed, red line) is aligned with the envelope of the original signal (solid, blue line). As the frames progress, the degraded envelope differs more and more from the original envelope, with two distinctive, artificial formants standing out in the last frame, shown in the bottom right graph. These formants differ in amplitude, frequency, and bandwidth from the ones in the original signal. Therefore, we can conclude that the degraded signal is impaired by an artifact. The fundamental frequency (around 230 Hz) is approximately the same in both signals, although the formants of the degraded signal do not align with the original signal. Because of this, we also used formant information, which we extracted from the spectral envelope of the original and the degraded signal as input features of the model. Additionally, the power of each frame was calculated as RMS (root mean square) and used as feature. All features were calculated for frames with a length of $t_{\text{frame}} = 60\,\text{ms}$ and an overlap of 70%, yielding a frame every 18 ms. Thus, we avoid being in sync with the original 20 ms frames used by the codec, since in practice this information wouldn't be available either.

2.1. Fundamental frequency

To capture the effect of a constant fundamental frequency, caused by repeating vocal tract information from the last correctly received frame, we firstly downsampled the speech signals to $f_s = 16\,\text{kHz}$ and then applied a low-pass filter with cut-off frequency $f_c = 1500\,\text{Hz}$. The autocorrelation function $r_{yy}(k)$ is computed by using the FFT/IFFT method and all values are normalized to the autocorrelation value at zero lag $r_{yy}(k) = r_{yy}(k)/r_{yy}(0)$, where k is the lag in samples. Then, all values before the first zero crossing are set to zero and the autocorrelations of all frames are combined to a matrix $\mathbf{R}_{yy}$, where $l \in \{1, ..., N_l\}$ is the frame index. The resulting matrix can also be called *autocorrelogram* and is shown in Figure 1.

To extract the $f_{0,y}$ ridge, firstly we find the location of the maximum value in the autocorrelogram of the degraded speech file $f_{0,y}(l)$. Then we search left of the maximum to find the maximum value in the neighbouring 11 lag bins $R_{yy}(k_{\text{s},y}, l-1)$ with $k_{\text{s},y} \in \{f_{0,y}(l)-5, ..., f_{0,y}(l)+5\}$. This is repeated until the maximum in the neighbouring bins is below a threshold of $\lambda_{0,y} = 0.6$. After that, we repeat the same procedure on the right side of the initially found maximum. As a next step, we search for the corresponding $f_{0,x}$ ridge in the original speech file. To this end, the maximum in the reference autocorrelogram is found, only considering the frames in which we found the $f_{0,y}$ ridge in the degraded file. Again, we search to the left of the maximum and then to the right with a lower threshold of $\lambda_{0,x} = 0.4$. However, to avoid obtaining divergent fundamental frequencies in the speech signals, only due to minor periodicity differences in the autocorrelation functions, a weighting function is applied that favours values along the degraded ridge $f_{0,y}$. This can be written as

$$f_{0,x}(l) = \arg\max_{k_{\text{s},x}} \left(\text{R}_{xx}(k_{\text{s},x}, l) - \alpha |(k_{\text{s},x} - f_{0,y}(l)| \right), \quad (1)$$

where $\alpha = 0.035$ describes how strong the influence of the weighting is. When both ridges are found, all values in the frames where the ridges were found are set to zero, and the next ridge is extracted by finding the maximum value in $\mathbf{R}_{yy}$. This is repeated until no values greater than $\lambda_{0,x} = 0.7$ are left in $\mathbf{R}_{yy}$. From both f_0 ridges we calculate a fundamental frequency distance, measured in sample lags, as follows:

$$f_{0,d}(l) = f_{0,y}(l) - f_{0,x}(l). \quad (2)$$

As a further feature we extract the difference along the ridge $f_{0,y}(l)$ of both autocorrelograms as follows:

$$p_{d,y}(l) = \mathbf{R}_{yy}\left(f_{0,y}(l), l\right) - \mathbf{R}_{xx}\left(f_{0,y}(l), l\right). \quad (3)$$

This feature tells us how large the difference in periodicity of both signals is. If there is a strong periodicity in the degraded signal, which is missing in the original signal, we can assume that an artificial, periodic sound has been added to the degraded signal by the concealment algorithm. On the contrary, if a periodic sound is missing in the degraded file, we can assume that parts of a vowel are missing in the speech signal. Therefore, we calculate the difference along the ridge of the reference signal $f_{0,x}(l)$ as well:

$$p_{d,x}(l) = \mathbf{R}_{xx}\left(f_{0,x}(l), l\right) - \mathbf{R}_{yy}\left(f_{0,x}(l), l\right). \quad (4)$$

In case there was no ridge found in a frame l, the distances are set to zero $f_{0,d}(l) = p_{d,y}(l) = p_{d,x}(l) = 0$. As a result, we receive one value for each feature every $18\,\mathrm{ms}$.

2.2. Formants

To extract the formant features, the signals are firstly resampled to $f_s = 14\,\mathrm{kHz}$. After that, the coefficients of a 14th order, forward linear prediction are calculated for each speech frame. The frequency response of the coefficients then gives the spectral envelope of the original signal $S_x(l)$ and the degraded signal $S_y(l)$, where l is the frame index. The resulting envelopes are used to find the formants by peak extraction. For each peak in the envelope, the amplitude $f_{\mathrm{amp}}(b,l)$, frequency $f_{\mathrm{frq}}(b,l)$, prominence $f_{\mathrm{prm}}(b,l)$, and the width $f_{\mathrm{wid}}(b,l)$ are computed [11]. The first four peaks with a prominence higher than $\lambda_{prom} = 0.5$ are then kept, with $b \in \{1,2,3,4\}$. From these formants the distance between the original and degraded signal in amplitude, frequency, prominence and width is calculated for each formant found in the original envelope and also for each formant found in the degraded envelope. The corresponding formant that is used for comparison is obtained by finding the formant with the closest frequency to the frequency of the considered formant, as follows:

$$b_{1,\mathrm{cmp,x}} = \arg\min_{b}\left(f_{\mathrm{frq},x}(1,l) - f_{\mathrm{frq},y}(b,l)\right), \quad (5)$$

where $f_{\mathrm{frq},x}(1,l)$ is the considered formant from the original signal, for which a distance is measured. The index of the comparison formant b_{cmp} is then also calculated for the other 3 peaks $b = \{2,3,4\}$ and for the degraded signal. Subsequently, the distances are calculated as follows:

$$f_{\mathrm{amp},xy}(1,l) = f_{\mathrm{amp},x}(1,l) - f_{\mathrm{amp},y}((b_{1,\mathrm{cmp,x}}),l) \quad (6)$$

Again this is calculated vice versa for the degraded formants and also for the frequency, prominence, and width. These distances are then used together with the formant frequency, prominence, and width as feature inputs, yielding in total 14 formant feature vectors with four entries each.

3. Decision Trees

The extracted features from the autocorrelogram and the spectral envelope are then used to train a binary classifier that distinguishes between unimpaired frames and frames that contain packet-loss. The features that we calculated are often linked to each other and only give a meaningful statement when considered together. For example, a frame is likely to contain packet-loss if the fundamental frequency differs by 100 Hz, but only if the amplitude of the first formant differs by at least 20 dB as well. These kind of features are suitable to be trained with a binary decision tree [12]. Additionally, decision trees allow for an easy interpretation and hence help to find possible improvements of a model.

3.1. Packet loss detection

In a first step, we train a decision tree to find frames that contain packet-loss. To this end, we used the packet-loss patterns that we generated to simulate packet-loss as response variable. With these patterns we know exactly in which frames the decoder had no speech data available and consequently applied a concealment algorithm.

3.2. Perceived packet loss detection

After the frames with the assumed packet-loss were found, we applied a second decision tree that classifies whether the found packet-loss was perceivable as a quality degradation or not. To do this, we used annotations from an expert listening experiment and applied them as response variable in the decision tree learning process. The motivation behind this, is on one hand to also find frames affecty by packet-loss that can be detected in the signal, but in fact don't make an impact as a quality impairment. On the other hand, we anticipate a more accurate classification when the decision tree can concentrate on potential packet-loss candidates only. In the training phase of this second stage, we applied light additive white noise before calculating the input features. Thus, rendering the model more robust towards noisy conditions.

4. Databases

We used two databases for training, and one database for testing. The degraded speech signals were generated by randomly selecting a bitrate mode of the respective codec (either EVS or AMR-WB). DTX was set off for all conditions. Then, exactly one burst of lost packets, at a random location, was applied to the speech file with the ITU-T EID (error insertion device) tool [13]. The length of each burst was a random number of consecutive lost frames between 1-6. Both codecs use a standard frame length of 20 ms, resulting in lost speech length from 20 ms to 120 ms. Where we assume that lost speech of a duration of 20 ms is not necessarily perceivable as a degradation, but 120 ms of lost speech is almost certainly perceivable as a quality impairment, when the erased packets were located within an active speech segment of the file.

The files were then annotated in an expert listening test on a four point category-ratio scale [14], which is presented in Figure 3. The test participants were experts who are dealing with speech processing as part of their work. During the test it was possible for the experts to listen to a condition as often as they needed to make a confident decision. In these experiments we are not interested in a subjective opinion of the participant, but rather want to find out if the packet-loss is perceivable at all.

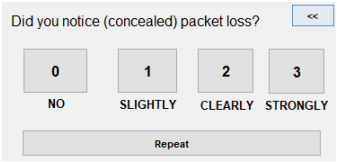

Figure 3: *Four point CR scale of expert listening test.*

A - Training: The first database was generated from four German and four English sentences taken from Annex C of ITU-T Rec. P.501. These speech files are specifically prepared for use with speech quality prediction models [15]. Based on those files, we generated 500 degraded speech files that were

coded with EVS and contain one packet-loss burst each. The files were then annotated by one expert in an auditory experiment.

B - Training: This databases doesn't contain any packet-loss, its purpose is merely to make the detection model more robust towards different speaker characteristics and low bitrate coding, which can be confused with packet-loss concealment. We used 100 German semi-spontaneous dialogues from the NCS corpus [16] that we coded with random bitrate modes.

C - Test: This database was generated from 100 English double sentences that were used during the ITU-T P.863 [2] selection phase. From these files the individual sentences were extracted, omitting the speech pause before and after, thus resulting in 200 speech files. Three expert listeners annotated this database in an auditory test, one of which also annotated database A. An overview of the databases can be seen in Table

Table 1: *Overview of databases.*

	#	EVS / AMR-WB	PL	Dur.	Tot. dur.	S.	L.	M / W	f_s
A	500	500 / 0	y	8 s	66 min	4	1	50 %	48 kHz
B	100	50 / 50	n	30-90 s	71 min	100	0	50 %	48 kHz
C	200	100 / 100	y	2-3 s	8 min	4	3	50 %	48 kHz

1, where "L." is the number of listeners and "S." is the numbers of speakers. Note that for the AMR-WB conditions we firstly downsampled the speech files to 16 kHz and then upsampled the files to 48 kHz again after decoding.

5. Results and discussion

The features were calculated as described in the sections before and used to train the decision trees. Only lost packets that were at least "clearly" perceived were used in the training phase of the second decision tree. In the training phase of the first decision tree, the generated error patterns were used as response variable. Then we applied the models to the test database C. We evaluated the EVS and AMR-WB conditions of database C separately, since no AMR-WB transmission errors were used in the training phase. The training database B was omitted in the evaluation because it doesn't contain any transmission errors and was only used to increase the robustness of the model.

To evaluate the detection model we concentrated on four evaluation metrics. Firstly, we calculated the number of correctly found lost packets that were at least "clearly" perceived by the experts. These would be all lost packets with a mean rating of 1.5 or higher. Because the packet-loss concealment algorithm may influence the speech signal even after the location of the actually lost packet, we used a tolerance of 200 ms for the evaluation. If at least one of the frames within the tolerance was detected as packet loss, it was counted as detected and true positive (this means the maximum number of true positives per condition is one). If a frame was detected as packet-loss outside the tolerance it was counted as false positive detection. The *TPR-CP (true positive rate - clearly perceived)* is then the number of clearly perceived true positives, divided by the total number of conditions with clearly perceived packets. The *FPR (false positive rate)* is the number of frames, falsely detected as packet-loss, divided by the number of conditions. Note that we did not divide by the number of unimpaired frames since almost all frames are unimpaired and only few contain packet-loss concealment. The *TPR-SP (true positive rate - slightly perceived)* is the number of slightly perceived true positives, divided by the total number of conditions with slightly perceived packets. The *FPR-NP (false positive rate - not perceived)* is the number of detected packet loss conditions that were not perceived, divided by the total number of conditions with not perceived packet-loss.

Table 2 presents the results after the first decision tree classifier. Most of the "clearly" perceived lost packets were found with a true positive rate of higher than 0.93. In the test database A there were zero falsely detected lost packets. In the test database C with EVS conditions, the false positive rate was relatively high with 0.05. The "slightly" perceived lost packets were detected with a ratio between 0.63 and 0.8. Furthermore, 22% - 43% of the lost packets that were not perceived were detected by the first tree. The second decision tree used the annotations

Table 2: *Results after first packet-loss decision tree.*

	TPR-CP	FPR	TPR-SP	FPR-NP
A - EVS (Train)	0.93	0.00	0.63	0.22
C - EVS (Test)	0.96	0.05	0.66	0.33
C - AMRWB (Test)	0.96	0.02	0.80	0.43

Table 3: *Results after second perceived packet-loss tree.*

	TPR-CP	FPR	TPR-SP	FPR-NP
A - EVS (Train)	0.91	0.00	0.43	0.11
C - EVS (Test)	0.96	0.00	0.62	0.22
C - AMRWB (Test)	0.89	0.02	0.80	0.20

from the expert listening experiment in the training phase and detects whether a lost packet was perceivable or not. The results after this second tree are shown in Table 3. The detection rate of the lost packets that were not perceived is now dropped by approximately 45%, whereas the TPR-CP rate of "clearly" perceived detected lost packets did not change significantly. Only the detection rate of the AMR-WB conditions of database C decreased. However, in the training phase of both trees no AMR-WB packet-loss conditions were included. Considering this, the performance on the AMR-WB conditions is still decent and proves that our approach is mostly independent of the applied codec. The FPR of the EVS conditions in database C is now zero, this means that the detector did not falsely find any packet-loss in unimpaired speech frames of EVS conditions. We can conclude from the good results after the first decision tree that auditory experiments are not necessarily needed to train the detector. This shows that it is possible to train the model with large amounts of data, generated with random error patterns.

6. Conclusions

In this paper, we showed that the approach of using formant information and fundamental frequency works well for the purpose of finding packet-loss, with consistent detection rates of 0.89-0.96 and almost no false positive detection. The results are a great step forward for the work item P.TCA. Furthermore, it was found that the approach works independently of the applied codec and no auditory annotations are necessary to train a first classifier. Thus, in future work we plan to use larger databases for training, and to extend this model to more realistic scenarios, including live recordings with other distortions and jitter buffer. This will make a more advanced preprocessing and time alignment necessary. Also, a more robust calculation of the formants could improve the detection accuracy. Why certain packet-loss frames that are not perceived are detected in the signal and others are not detected needs to be investigated in further studies.

7. References

[1] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," 1996.

[2] ITU-T Rec. P863, "Perceptual objective listening quality assessment," 2014.

[3] M. Wältermann, *Dimension-based Quality Modeling of Transmitted Speech*. Berlin, Heidelberg: Springer, 2012.

[4] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Berlin, Heidelberg: Springer, 2011.

[5] ITU-T SG12 TD 137, "Technical requirement specification P.AMD and P.SAMD," 2017.

[6] ITU-T SG12 TD 122, "Requirement specifications for P.TCA (technical cause analysis)," 2017.

[7] J. Lecomte, T. Vaillancourt, S. Bruhn, H. Sung, K. Peng, K. Kikuiri, B. Wang, S. Subasingha, and J. Faure, "Packet-loss concealment technology advances in EVS," pp. 5708–5712, April 2015.

[8] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications," Geneva, 2004.

[9] 3GPP TS 26.441, "Codec for Enhanced Voice Services (EVS); General overview."

[10] 3GPP TS 26.171, "Speech codec speech processing functions; adaptive multi-rate - wideband (AMR-WB) speech codec; general description."

[11] MathWorks. (2018, Mar) Matlab 2017b documentation. [Online]. Available: https://mathworks.com/help/signal/ref/findpeaks.html

[12] L. Breiman, J. Friedman, and R. Olshen, *Classification and regression trees*, ser. The Wadsworth statistics, probability series. Belmont, Calif.: Wadsworth Internat. Group, 1984.

[13] ITU-T Rec. G.191, "ITU-T software tool library 2009 user's manual," 2009.

[14] S. Möller, *Quality Engineering. Qualität kommunikationstechnischer Systeme.*, 2nd ed. Berlin, Heidelberg: Springer, 2017.

[15] ITU-T Rec. P.501, "Test signals for use in telephonometry," 2017.

[16] L. Fernández Gallardo and B. Weiss, "The nautilus speaker characterization corpus: Speech recordings and labels of speaker characteristics and voice descriptions," in *Proc. of International Conference on Language Resources and Evaluation (LREC)*, 2018.