Urgent Voicemail Detection Focused on Long-term Temporal Variation

Hosana Kamiyama, Atsushi Ando, Ryo Masumura, Satoshi Kobashikawa and Yushi Aono
NTT Corporation, NTT Media Intelligence Laboratories, Japan
E-mail: hosana.kamiyama.tf@hco.ntt.co.jp Tel: +81-46-859-4624

***Abstract*—This paper proposes a effective urgent speech detection for voicemails focused on speech rhythm. Previous techniques use short-term features with millisecond scale (such as fundamental frequency, loudness and spectral features), and conventional techniques for urgent speech detection use also features obtained from entire speech (such as average speech rate). However, the features obtained from entire speech are too over-smoothed to explain the difference between urgent and non-urgent speech. We found that there was a difference between urgent and non-urgent speech in temporal variability related to speech rhythm. To handle the temporal variability of speech rhythm, the proposal extracts long-term temporal features. The long-term temporal features are envelope modulation spectrum and temporal statistics of Mel-frequency cepstrum coefficient with 1 sec scale. To use both features with different time scales, the proposed method integrates the long-term temporal features and the short-term features on neural networks. Our proposal yields better accuracy than the conventional methods (which uses e features obtained from entire speech); it achieves a 50.0% reduction in the error rate.**

I. Introduction

Some contact centers provide automatic reception services based on voicemail systems during non-business hours. It is important that contact centers preferentially respond to urgent voicemail messages. There some urgent voicemails that cannot be judged only by the text information obtained by automatic speech recognition. Therefore, the aim of this paper is to detect urgent voicemails without using text.

Techniques have been proposed for estimating paralinguistic information, such as emotion and urgency. The methods of [1], [2] use prosodic features, such as fundamental frequencies (F_0), loudness, and average speech rate. The methods focus on the difference in average prosody between urgent and non-urgent speech; the former have higher F_0, loudness, and speech rate. In tasks other than voicemails, such as 911 calls [3], [4], smart speaker [5], [6] and shout detection [7], the urgent situation detection techniques use spectral features, such as Mel-frequency Cepstrum Coefficients (MFCC), etc. The MFCCs represent the vocal tract characteristics of speech. The methods focus on the differences in the vocal tract characteristics between urgent and non-urgent situations even for the same word (e.g. "Help"). In emotion recognition tasks, features other than F_0, power and MFCC have been used to improve estimation accuracy [8], [9], [10], [11], [12]. Most emotion recognition techniques use F_0, loudness and MFCC as are used for urgency detection. As evidenced by the emotion recognition schemes that use speech rate [13], speech rate is an effective feature for urgency detection. To use speech rate, it is thought necessary to transcribe speech manually or recognize speech by using automatic speech recognition (ASR).

The more urgent the voicemail is, the more variable its speech rhythm often becomes; the speech rhythm in urgent situations often is disturbed. The previous methods do not handle the variability of speech rhythm. The features that have millisecond scales, which are called low-level descriptors (LLDs) and used for emotion recognition, can capture rapid and local variability of prosodic and spectral characteristics, but cannot model speech rhythm on the time scale of seconds as their temporal features are too short-term. On the other hand, the features obtained from entire voicemails such as average speech, which are used in previous techniques for urgent voicemail detection, rate are over-smoothed and so cannot explain the variability of speech rhythm. To detect urgent voicemails more accurately, it is necessary to use temporal features that can well represent the variability of speech rhythm.

This paper proposes urgency detection techniques focused on the temporal variability of speech rhythm. The proposed method uses not only short-term features on the millisecond scale, but also long-term temporal features on the second scale to well handle the temporal variability of speech rhythm. The long-term temporal features are introduced to capture the variability that short-term features have difficulty in modeling and that can be over-smoothed by features of the entire voicemail. As long-term temporal features, this paper uses envelope modulation spectrum (EMS) and temporal statistics of MFCCs. EMS is a representation of the slow amplitude modulation in a signal and is used for discriminating dysarthria [14]. The temporal statistics of MFCCs is a representation of the temporal phonetic variation and is used for speech rate estimation [15]. To well integrate the temporal features that have different periods, the proposed method inputs the features to different recurrent neural networks (RNNs). The features are concatenated in the neural networks including the RNNs, and used for urgent voicemail detection. We conduct an experiment to rank our proposals against previous techniques.

II. Urgency detection

This section describes the baseline and the conventional methods of urgency detection.

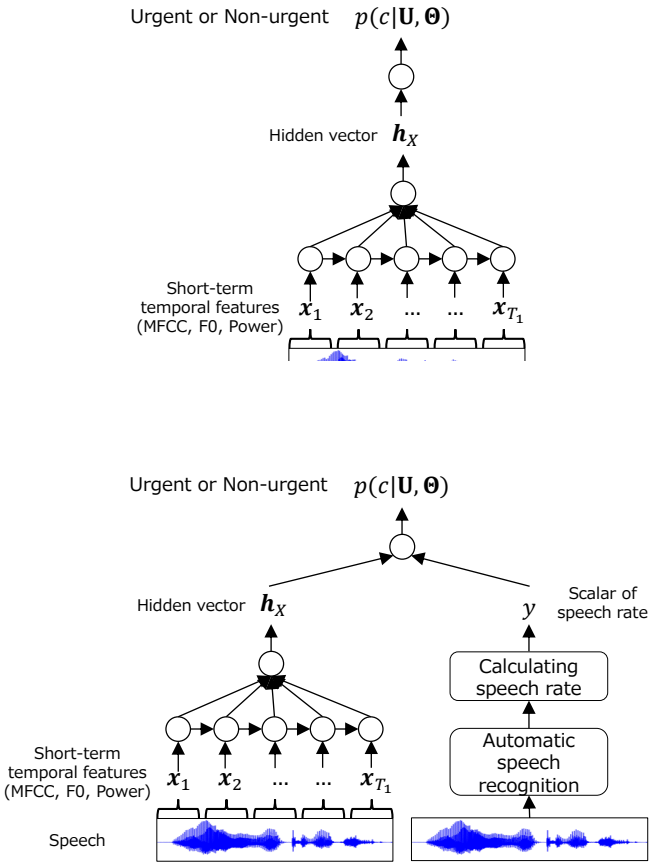

Fig. 2. Conventional neural network

A. Task description

The task of this paper is detection of voicemail urgency. The classes are *urgent* and *non-urgent*. *urgent* means a voicemail whose creator (customer) requires an urgent response, *non-urgent* means a voicemail that does not require such a response. Estimation of urgent or non-urgent voicemail can be formulated as follows:

$$\hat{c} = \arg\max_{c \in C} p(c|\mathbf{U}, \boldsymbol{\Theta}) \tag{1}$$

where $\hat{c}$ is estimated urgent, C is the set of urgency classes where we set $C = \{0, 1\}$ (Class 1 means *urgent* and Class 0 means *non-urgent*), $\mathbf{U}$ represents the features of the voicemail and $p(c|\mathbf{X}, \boldsymbol{\Theta})$ is a urgency estimation model that outputs the posterior probability of urgency y when given $\mathbf{U}$ and $\boldsymbol{\Theta}$. In this paper, the feature set $\mathbf{U}$ and the model architecture of $p(c|\mathbf{U}, \boldsymbol{\Theta})$ are different between the baseline, conventional and proposed methods. The models are based on neural networks.

B. Baseline method

The baseline uses only low-level descriptors (LLDs) which are short-term features. The model architecture is shown in Fig. 1. The model is based on emotion recognition techniques [9], [11]. The network architecture is as follows:

$$\mathbf{h}_X = \texttt{RNN}(\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}; \boldsymbol{\Theta}_X) \tag{2}$$

$$p(c|\mathbf{U}, \boldsymbol{\Theta}) = \texttt{SOFTMAX}(\mathbf{w}_1 \mathbf{h}_X + \mathbf{b}_1) \tag{3}$$

where `RNN()` is a recurrent neural network which takes as input LLDs $\mathbf{x}_1, \cdots, \mathbf{x}_{T_1}$, and outputs hidden vector $\mathbf{h}_x$, $\boldsymbol{\Theta}_X$ represents the trainable parameters of `RNN`, `SOFTMAX()` is a function to calculate softmax activation in (), $\mathbf{w}_1$ and $\mathbf{b}_1$ are trainable parameters to perform a linear transformation of $\mathbf{h}_X$. When multiple features are used, such as MFCC, F_0 and power, the LLDs are analyzed using the same analysis window length. The network does not use average speech rate for urgency detection.

C. Conventional method

The conventional techniques uses average speech rate, which is a long-term feature not a temporal feature, for urgency detection [2]. The model that combines LLDs and the metric (scaler) of average speech rate for urgency detection, is shown in Fig. 2. To use average speech rate features, the network architecture is as follows:

$$\mathbf{h}_X = \texttt{RNN}(\mathbf{x}_1, \cdots, \mathbf{x}_T; \boldsymbol{\Theta}_X) \tag{4}$$

$$p(c|\mathbf{U}_X, \boldsymbol{\Theta}) = \texttt{SOFTMAX}(\mathbf{w}_2[\mathbf{h}_X, y] + \mathbf{b}_2) \tag{5}$$

where $[\mathbf{h}_X, y]$ is the concatenation of vector $\mathbf{h}_X$ and scaler y, $\mathbf{w}_2$ and $\mathbf{b}_2$ are a trainable parameters that allow linear transformation of the concatenated vector $[\mathbf{h}_X, y]$. To calculate average speech rate y, ASR is used for counting mora. The speech rate can be calculated by dividing the number of mora counted by ASR by the duration of the speech.

III. Proposed methods

This section describes the proposed methods for urgency detection focusing on speech rhythm.

A. Approach

Our approach uses the long-term temporal features to obtain the variability of speech rhythm. We assume that the more urgently a voicemail is uttered, the more variable its speech rhythm becomes; the rhythm of speech uttered in urgency situations is often disturbed. To handle speech rhythm variability, EMS and temporal statistical of MFCCs are used for urgent voicemail detection. EMS and temporal statistics of MFCCs are long-term temporal features with scale of 1 second, and are related to speech rhythm according to [15]. Fig. 3 and Fig. 4 show urgent and non-urgent speech signals and the features of the same speaker, respectively. To plot the fluctuation in speech features, principal component analysis (PCA) of the speech rhythm features was performed, and the first principal component of result of PCA are shown. It is clear that the PCA of EMS of non-urgent voicemails exhibit more regular fluctuation than urgent speech and the PCA of statistical of MFCCs demonstrate less fluctuation than urgent speech. Our proposal uses this difference in feature fluctuation between urgent and non-urgent speech.

B. Long-term temporal feature

This paper uses long-term temporal features based on the online speech rate estimation technique [15]. We use two speech features: envelope modulation spectrum (EMS) and

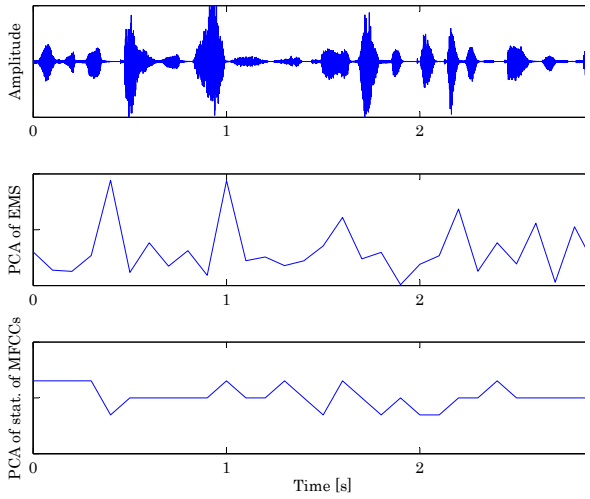

Fig. 3. Urgent speech and the first principal component of feature related to speech rhythm

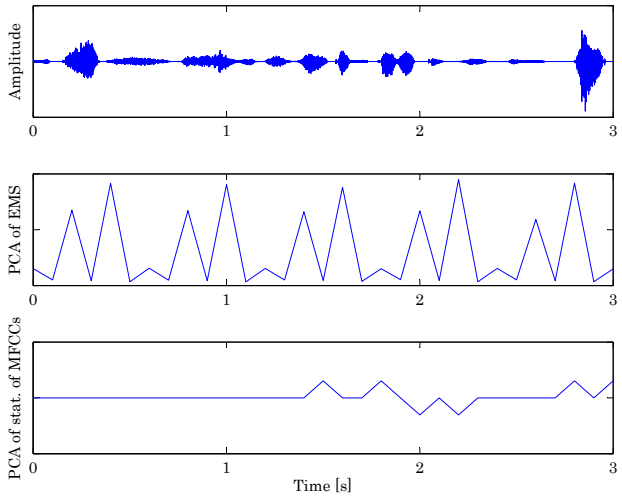

Fig. 4. Non-urgent speech and the first principal component of feature related to speech rhythm

temporal statistics of MFCCs. To obtain speech features, the analysis window was set longer than that used in the extraction of LLDs.

1) EMS: EMS is used for discriminating dysarthria as speech rhythm features, and is a representation of slow amplitude modulation in a signal [14]. The speech signal is passed through a range of octave band filters, after which an envelope is extracted from each individual octave. EMS is feature based on characteristics of envelope, and consists of peak frequency, peak amplitude, and energy ratio of spectrum of envelope.

2) Temporal statistics of MFCCs: The temporal statistics of MFCCs, which are representations of phonetic variation, are calculated from the LLD of MFCC [15]. After extracting LLD, the statistics of LLD are calculated on the second scale window, such as 1 second. As the statistics, we calculate the mean, standard deviation, maximum value, skewness, kurtosis, and mean absolute deviation within the long frame window according to [15].

C. Fusion of short-term features and long-term temporal features in neural network

Our proposed network inputs the features gathered using different analysis window lengths. Short-term features capture rapid and local variability of LLDs. Our proposed long-term features handle the slow variation in speech rhythms. The model architecture, shown in Fig. 5, is as follows:

$$\mathbf{h}_X = \texttt{RNN}(\mathbf{x}_1, \cdots, \mathbf{x}_T; \mathbf{\Theta}_X) \quad (6)$$

$$\mathbf{h}_Z = \texttt{RNN}(\mathbf{z}_1, \cdots, \mathbf{z}_{T_2}; \mathbf{\Theta}_Z) \quad (7)$$

$$p(c|\mathbf{U}, \mathbf{\Theta}) = \texttt{SOFTMAX}(\mathbf{w}_3[\mathbf{h}_X, \mathbf{h}_Z] + \mathbf{b}_3) \quad (8)$$

where $\mathbf{z}_1, \cdots, \mathbf{z}_{T_2}$ are the speech rhythm features, $\mathbf{h}_z$ is hidden vector obtained by RNN, $\mathbf{\Theta}_Z$ is the trainable parameters of RNN, $\mathbf{w}_3$ and $\mathbf{b}_3$ are trainable parameters needed to perform linear transformation of concatenated vector $[\mathbf{h}_X, \mathbf{h}_Z]$. To input the two features, the network has two input layers with different analysis window lengths.

IV. Experiments

To evaluate the proposed methods, we conducted urgency detection using simulated voicemails data.

A. Dataset

In this paper, we use newly recorded and annotated voicemails. The task is frozen food selling and includes several subtasks such as inquiries, cancelling or additional orders during non-business hours. First, we set the situation, and urgency information in each sub-task to create 12 useful scenarios. Next, we recorded voicemails via phone sets while following the scenarios but the speech sentences was uttered on the fly. The result was 240 voicemails by 20 speakers. Total length was 2.2 hours and the average length was about 30 seconds. All were monaural recorded, 8 kHz with 16 bit format.

Urgency labels for voicemails were annotated by three people. Samples were used for urgent voicemail detection only if all annotators assigned the same urgency labels to each. Finally, we obtained 120 urgent samples and 100 non-urgent samples. Cohen's kappa of labeling urgency for three annotators was 0.89.

B. Setup

1) Feature extraction: We extracted three features with different windows: 1) short-term features called LLDs with millisecond time scale (Short-term), 2) long-term temporal features with second time scale (Long-term temporal), 3) features extracted from entire voicemails (Entire). The short-term features were 28 dimensional acoustic features; 12 dimensional Mel-Frequent Cepstral Coefficients (MFCCs), loudness, fundamental frequency (F_0), the first order derivatives of MFCCs , loudness, and F_0. The features were extracted using 25 millisecond windows with 10 millisecond shift. The long-term temporal features were envelope modulation spectrum (EMS) and temporal statistics of MFCCs. The features were

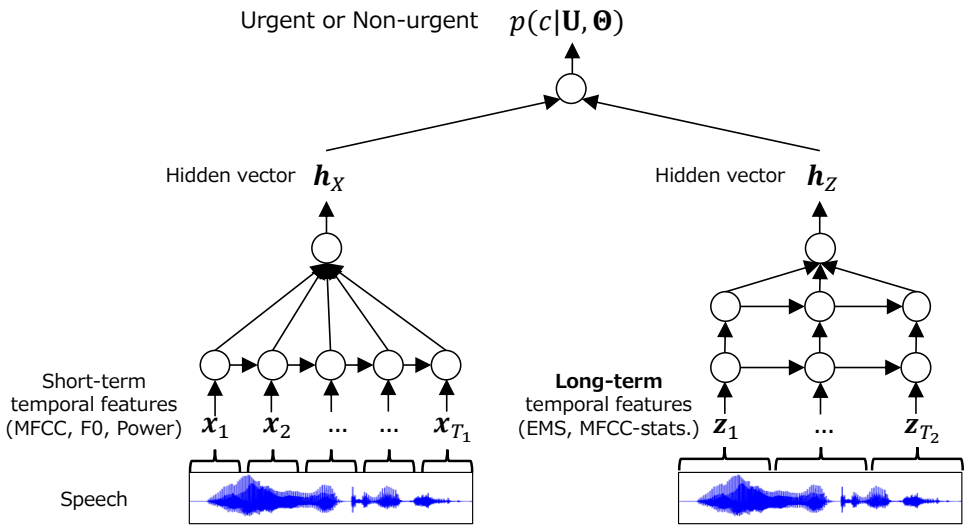

Fig. 5. Proposed neural network

TABLE I
EXTRACTION CONDITIONS OF ENVELOPE MODULATION SPECTRUM (EMS)

Center frequency of octave band [Hz]	30, 60, 120, 240, 480, 1920, 3480
Feature metrics	Peak frequency in 0-10Hz Peak amplitude in 0-10Hz Energy from 3-6Hz Energy from 0-4Hz Energy from 4-10Hz Energy ratio between 0-4 Hz and 4-10 Hz

extracted using 1 second windows with 10 millisecond shift. The conditions of EMS extraction are shown in Table I. EMS dimension was 42, which are based on a 7 octave band filter and 6 feature metrics. The six statistical features of MFCCs consisted of mean, standard deviation, maximum value, skewness, kurtosis and mean absolute deviation of MFCCs (extracted as 12 dimensional MFCCs) and power, the first, second order derivatives of them. The dimension of MFCC statistics was 39 * 6 = 234. The long-term (but not temporal) features was average speech rate of each entire voicemail. The average speech rate was calculated by two methods: 1) automatic determination by automatic speech recognition (ASR), 2) manual calculation by using transcribed text (Oracle).

2) Neural network training: In the baseline, the conventional and the proposed network, LSTM with attention layer was used for RNN; the output of which was a 32 dimensional hidden vector. In the proposed network, two LSTMs were used for RNN which outputs a 64 dimensional hidden vector. Optimizer was Adam with the learning rate of 0.001; the dropout rate was 50%. We measured performance in term of accuracy as determined by 10-fold cross validation. We used 1/10 of the samples as the test set. We trained all models 5 times in each condition and compared the accuracy attained relative to the highest performance.

C. Results and discussion

Performance comparisons of the baseline, the conventional and the proposed method are shown in Table II. Our proposal, which used both short- and long-term temporal features at no.(9), yielded the highest performance, and achieved 89.5% and 50.0% error reduction from the conventional method using oracle average speech rate (3). This shows that the long-term temporal features related to speech rhythm are more effective for urgency detection than the speech rate features of voicemails.

The conventional methods, (2),(3), are better than the baseline method, (1) with regard to average speech rate. The proposed methods, (7), (8), (9), are superior to conventional methods, (2), (3). It is indicated that our proposed long-term temporal features explain the difference between urgent and non-urgent speech more clearly than the average speech. In the proposed methods, the short-term features, (7), (8), (9) are also more effective than using only EMS and the MFCC statistics, (4), (5), (6). This indicates that the short- and long-term features focus on different features of urgent speech. In urgent situations, the short-term features may be intended to implement rapid changes in prosody and voice tract, while the long-term features may be intended to handle slow changes in prosody, such as speech rhythm. In the proposed methods, the temporal statistics of MFCCs, (5), (8) are more effective than EMS, (4),(7). Since Japanese is a pitch accent language, the EMS did not represent more than the temporal statistics of MFCCs; EMS is related to speech rhythm as altered by stress but not pitch accent. The EMS might be effective for urgency detection if stress accenting is prevalent.

V. CONCLUSIONS

In this paper, we proposed an urgent message detection method that uses long-term temporal features. The previous methods use short-term features, called LLDs, and average

TABLE II
ACCURACY OF URGENCY DETECTION.

		Short-term	Long-term temporal		Entire	Acc.
			EMS	MFCC	Speech rate	
(1)	Baseline	✓	-	-	-	.748
(2)	Conventional	✓	-	-	✓(ASR)	.776
(3)		✓	-	-	✓(Oracle)	.790
(4)	Proposed	-	✓	-	-	.709
(5)	(no short-term)	-	-	✓	-	.732
(6)		-	✓	✓	-	.755
(7)	Proposed	✓	✓	-	-	.791
(8)		✓	-	✓	-	.886
(9)		✓	✓	✓	-	.895

speech rate of entire voicemails. However, the features obtained from entire voicemails cannot explain the difference in temporal speech rhythm between urgent and non-urgent voicemails. Our proposal utilizes the fact that the speech rhythm of urgent voicemails is more variable than that of non-urgent voicemails. To obtain the features related to speech rhythm, we extract the EMS and temporal statistics of MFCCs using windows of the scale of seconds. To handle different duration windows, the proposed network inputs the short- and the long-term temporal features into different RNNs, and uses the hidden vectors output by RNNs for urgency detection. Our proposed method achieved 89.5% higher accuracy and 50.0% stronger error reduction compared to the conventional method.

REFERENCES

[1] M. Ringel and J. Hirschberg, "Automated message prioritization: making voicemail retrieval more efficient," in *Proc. CHI*, 2002, pp. 592–593.

[2] Z. Inanoglu and R. Caneel, "Emotive aleart: HMM-based emotion detection in voicemail message," in *Proc. IUI*, 2005, pp. 251–253.

[3] I. Lefter, L. J. M. Rothkrantz, and D. A. Leeuwen, "Automatic stress detection in emergency(telephone) calls," in *International Journal of Intelligent Defence Support System*, 2016, pp. 148–168.

[4] R. M. Hegde, B. S. Manoj, B. D. Rao, and R. R. Rao, "Emotion detection from speech signals and its applications in supporting enhanced QoS in emergency response," in *Proc. ISCRAM*, 2006, pp. 82–91.

[5] E. Principi, S. Squatini, R. Bonfigli, G. Ferroni, and F. Piazza, "A speech-based system for in-home emergency detection and remote assistance," in *Journal of Expert System with Apprications*, vol. 42, 2015, pp. 568–5683.

[6] E. Principi, S. Squatini, R. Bonfigli, E. Cambria, and F. Piazza, "Acoustic template-matching for automatic emergency state detection: An ELM based algorithm," in *Journal of Neurocomputing*, vol. 149, 2015, pp. 426–434.

[7] J. Pohjalainen, P. Alku, and T. Kinnunen, "Shout detection in noise," in *Proc. ICASSP*, 2011, pp. 4968–4971.

[8] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proc. of Interspeech*, 2009, pp. 312–315.

[9] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using reccurent neural networks with local attention," in *Proc. of ICASSP*, 2017, pp. 2227–2231.

[10] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, and Y. Aono, "Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification," in *Proc. of ICASSP*, 2018, pp. 4964–4968.

[11] H. M. Fayek, M. Lech, , and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. IJCNN*, 2016, pp. 566–570.

[12] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. of ICASSP*, 2019, pp. 526–529.

[13] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, and Y. Aono, "Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center calls," in *Proc. Interspeech*, 2017, pp. 1716–1720.

[14] J. Liss, S. LeGendre, and A. Lotto, "Discriminating dysarthria type from envelope modulation spectra," in *Journal of Speech Language and Hearing Research*, 2010, pp. 1246–1255.

[15] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Online speaking rate estimation using recurrent neural networks," in *Proc. ICASSP*, 2016, pp. 5245–2237.