LSTM-BASED WHISPER DETECTION

Zeynab Raeesy, Kellen Gillespie, Chengyuan Ma, Thomas Drugman, Jiacheng Gu, Roland Maas, Ariya Rastrow, Björn Hoffmeister

Amazon Alexa

{raeesyzr, kelleng, mchengyu, drugman, jiacheg, rmaas, arastrow, bjornh}@amazon.com

ABSTRACT

This article presents a whisper speech detector in the far-field domain. The proposed system consists of a long-short term memory (LSTM) neural network trained on log-filterbank energy (LFBE) acoustic features. This model is trained and evaluated on recordings of human interactions with voice-controlled, far-field devices in whisper and normal phonation modes. We compare multiple inference approaches for utterance-level classification by examining trajectories of the LSTM posteriors. In addition, we engineer a set of features based on the signal characteristics inherent to whisper speech, and evaluate their effectiveness in further separating whisper from normal speech. A benchmarking of these features using multilayer perceptrons (MLP) and LSTMs suggests that the proposed features, in combination with LFBE features, can help us further improve our classifiers. We prove that, with enough data, the LSTM model is indeed as capable of learning whisper characteristics from LFBE features alone compared to a simpler MLP model that uses both LFBE and features engineered for separating whisper and normal speech. In addition, we prove that the LSTM classifiers accuracy can be further improved with the incorporation of the proposed engineered features.

Index Terms— whisper phonation, long-short term memory neural networks, whisper

1. INTRODUCTION

Advancements in speech and language technologies have resulted in the deployment of dialogue systems in real environments. These environments range from noisy living rooms with background noise to rooms that are very quiet. In the latter type of environment, a user may wish to whisper to the device, and in return would expect a response in a quieter and/or whispered voice. Triggering the suitable response mode on the device first requires detecting the whisper speech of the user. While automatic speech technologies are well researched and evaluated on normal speech in various acoustic conditions, there has been little effort in developing such technologies for other types of phonated speech such as whisper.

Whisper speech is mainly characterized as unvoiced speech due to a lack of periodic excitation in vocal folds. Studies of spectrograms have suggested that whisper speech overall has less energy at lower frequency bands compared to normal speech [1]. In [2] and [3], a consistent increase in F1 formant frequency in whisper speech in comparison to normal speech was reported. The signal characteristic differences between whisper and normal phonation have been the basis for several studies on classifying the two modes. Wenndt et. al [1] proposed a classification approach on whisper versus normal phonation using the energy ratios between high-frequency and low-frequency bands. In a study by Zhang and Hansen [4], several speaking modes such as whisper, soft, normal, loud, and shouted were investigated. A GMM-based classification system was developed based on sound intensity level, sentence duration, silence percentage, frame-energy distribution, and spectral tilt for these five categories of phonation. Spectral information entropy (SIE) was estimated in [5] from probability density functions calculated over all frequency components, and the energy ratios between SIEs of multiple frequency sub-bands were used to form a 9-dimensional feature vector used in training GMM-based vocal effect classifier. The insensitivity of these features to absolute energy values in turn makes the classifiers built with such features robust to varying energy levels in input signals. This work was further enhanced in [6] by using improved features with more frequency sub-bands for calculating SIEs.

To the best of our knowledge, the majority of the work in the related literature focuses on detecting and devising relevant features for classifying whisper and normal speech. Compared to traditional machine learning models, deep and recurrent neural network models are capable of learning complex features even from raw data, with less reliance on task specific engineered features. One such example is voice activity detection (VAD), where DNN-based approaches [7, 8] have proven superior in performance over classic approaches developed around engineered features [9, 10], such as audio energy [11], pitch [12], zero-crossing rate [13, 14], and cortical features [15].

In this paper, we propose using long-short term memory (LSTM) neural network models for the task of whisper detection. LSTM networks have proven to be successful classifiers in ASR, having been applied to a variety of tasks such as acoustic modeling [16] and endpoint detection [17]. We use a dataset of real recordings of natural human interactions, in

whisper and normal speech, with a far-field voice-controlled speaker. To the best of our knowledge, this is the first work reporting on application of deep neural networks to whisper speech detection. Comparing to a baseline simple multilayer perceptron, LSTMs are proven to achieve a significantly higher frame accuracy. Based on our observation of LSTM posterior trajectories, we examine a number of inference modules for classification of utterances. In addition, inspired by the literature on whisper classification, we study the application of engineered features to this task. We examine classification performance with simple multilayer perceptron (MLP) models and LSTMs by adding a 6-dimensional vector of engineered features that are useful in distinguishing whisper/normal speech and compare them with models trained only on LFBE features. Based on our findings we prove that, in addition to scaling better, a more complex model such as an LSTM can perform reasonably well without the computational burden of engineered features. Through experiments and evaluations, we show that LSTM's performance can be further improved using the additional engineered features.

This paper is organized as follows. In the first part, we present an overview of the proposed classifier and inference mechanisms in sections 2.1 and 2.2. Experiments and evaluations in section 3 comprises the dataset specifications in 3.1, metrics in 3.2, evaluation results in 3.3, and inference comparisons in 3.4. In the second part, sections 4.1 and 4.2 introduce the engineered features for whisper detection and evaluate the effectiveness of the features in classification task. Finally, the conclusions of this work and plans for future work are provided in section 5.

2. LSTM-BASED WHISPER DETECTOR

In this section, we begin with an overview of LSTM neural networks and their application to our whisper detector. We then discuss how we perform inference for utterance-level decision making from the frame-level posteriors of the LSTM classifier.

2.1. Overview of the Classifier

The input data to the whisper classifier is in the form of sequential frames. Standard feed-forward MLP networks, with no concept of memory, do not allow us to use this data in an intuitive sequential, contextual way. Recurrent neural networks (RNNs) use feedbacks from their internal states in processing sequences of inputs, and thus consider the history of their states when modeling sequential data. However, RNNs are limited to short-term memory, as they suffer from the vanishing/exploding gradient problem [18]. Long short-term memory (LSTM) models are extensions of RNNs, where memory cells with input, output, and forget gates are introduced at each recurrent layer to control the flow of information, consequently facilitating the learning of both short and long term dependencies in the input sequences [19].

For the whisper classifier, LSTM models are trained using sequences of frames and their labels. Since this application of the model requires utterance-level decisions, each utterance in our dataset is tagged as whisper/non-whisper. These tags are propagated as target labels to all frames of that particular utterance. The model is trained using a cross-entropy objective function and is optimized with stochastic gradient descent (SGD) [20] using the backpropagation through time (BPTT) algorithm [21].

2.2. Inference

The whisper classification models are structured to output scores at the frame level. Given a set of individual frame scores across a given utterance, we must then use an inference module, or result building process, to generate a classification score at the utterance level.

Upon performing posterior analysis of our model predictions on whisper (positive) test cases, we often observe sharp drops in posterior values towards the final frames of utterances. With the *last-frame* inference module, these drops in turn result in sudden changes in utterance level predictions. An example of this behavior is shown in Figure 1. After investigating the audio, we find that these drops generally coincided with short trails of silent or near-silent frames found at the end of utterances. In many of these cases, as shown by the aforementioned figure, the model is confident in predicting whisper for long periods of time, only to fall sharply in the final frames.

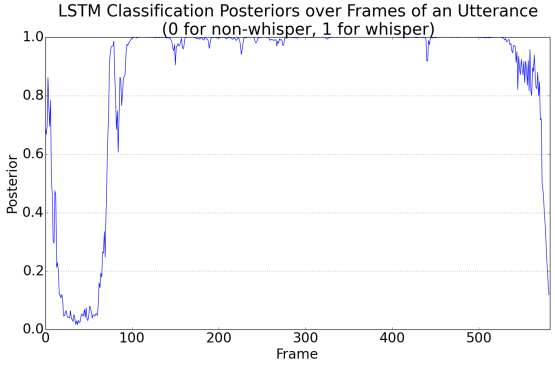

Fig. 1. LSTM posteriors over frames of an utterance

To better represent the model predictions over the course of an entire utterance, we experiment with simple alternative inference modules. There are three main 'classes' of inference module investigated, explained below.

last-frame: Takes the last frame posterior.

median: Takes the median posterior of all frames.

mean: Takes the mean posterior of all frames.

In addition, we investigate applying an offset to the computation windows of each inference module, append-

ing the module name with *ignore-last-50* in these cases. This offset of 50 frames is empirically derived from trailing silence lengths previously observed by the production end-of-utterance detector on live traffic data, and is utilized to explore whether the LSTM posteriors are more strictly tied to end of speech rather than end of the utterance.

Overall, we find that the inference modules that consider more frames outperform *last-frame*, notably in terms of recall. Windowed approaches using fewer frames were also considered, but proved difficult to tune and did not generalize well between different test sets. We also find that the offset only improves results for inference modules not considering many frames. The full results of the investigation are provided in section 3.4.

3. EXPERIMENTS AND EVALUATIONS

3.1. Data Preparation

A series of whisper and normal utterances in US English was recorded in a quiet environment setup. We refer to this test set as "in-house" in the rest of this paper. The microphones used for collection were each placed between 1–2 feet away from the speaker. With the exception of 3 recording sessions that had fan noise, the audio was generally recorded in clean conditions with no background noise. In addition, we incorporated data from real recordings of natural human interactions with voice-controlled far-field devices containing only normal speech. In both dataset categories, the audio sampling rate was 16kHz, and for each utterance, the target label (whisper or normal) was propagated to all frames in that utterance.

From the in-house dataset of roughly 28k utterances, around 23k utterances were used in training and cross-validation, and the remaining 5k were selected for testing purposes, with no speaker overlap between train and test sets. The in-house test set consists of 3670 whisper and 1565 normal utterances. From the real recordings set, we added 30k utterances for training and cross-validation, and withheld 11k utterances for evaluating the false-positive rate on normal speech.We refer to the former as "live traffic train" set and to the latter as "live traffic test" set.

3.2. Metrics

The ultimate goal of the whisper detector is to decide if an input speech signal is whisper or normal. Thus, in addition to raw frame accuracy accumulated from LSTM posteriors, we use recall and false-positive rate metrics at the utterance level. The frame accuracy is calculated at the default threshold of 0.5 with no tuning. To have meaningful comparison of the models, we tune the model thresholds to achieve 0.1% false-positive rate on in-house tests. The tuned operating point (OP) is then used to compare the models in terms of false-positive rate on live traffic test set and recall on in-house sets. To have the overall picture, we also compare the models in terms of F1 score on accumulation of test sets.

3.3. Classifier Evaluation

We extract 64-dimensional LFBE features for every 25ms frame of the utterance, with 10ms overlap between the frames. Channel mean subtraction (CMS) is applied to utterances on a per-speaker, per-device basis in real recordings and per-speaker in in-house test data. The LSTM model structure consists of 2 hidden layers each with 64 memory cells. The output layer is 2-dimensional, corresponding to whisper and normal status. The baseline system is a simple multilayer perceptron (MLP) with 3 hidden layers, each of size 40 units with ReLU activation, and a 2-dimensional output layer.

The final utterance level results are built using the mean of the frame posteriors of the entire utterance. This inference module was chosen empirically based on experiments explained in section 3.4.

Table 1 shows frame accuracy comparisons of the LSTM and MLP model on in-house and live traffic test sets. For fair comparison of models, the threshold has been tuned for both models to have equal false-positive rate on in-house test sets (the operating point OP). As observed, the LSTM outperforms the MLPs on both in-house recall and live traffic test sets FPR at false-positive rate of 0.1% on in-house. However a larger gap is observed on in-house test set which contains both whisper and non-whisper utterances, suggesting both positive and negative instances are contributing to the difference in the classification outcomes of the two models.

3.4. Inference at Utterance Level

Table 2 shows the results of our inference module investigation. As expected, the inference modules that consider more frames have improved recall and F1 scores over the *last-frame* module. The 50 frame offset only improves the performance of the *last-frame* module; it lowers the F1 score for the *median* module and performs worse in all metrics for the *mean* module. While the offset gives *last-frame* a higher probability of operating on a speech frame, as opposed to a frame after the end of speech, the trimming of the final 50 frames causes a loss of information in the modules considering all frames. Overall, the *mean* module with no offset had the best F1 score on this test set.

4. FEATURE STUDY

4.1. Classifier Features

Three categories of features are studied in this work: sum of residual harmonics (SRH), high-frequency energy (HFE), and features based on auto-correlation of time-domain signal (ACMAX). A review of these features is presented below.

Table 1. Comparison of LSTM and MLP trained on LFBE features on in-house and live traffic tests.

model	feature	in-house			live traffic tests	
		frame acc.	FPR (OP)	recall	frame acc.	FPR
MLP	LFBE	77.1%	0.1%	95.1%	94.9%	1.51%
LSTM	LFBE	93.5%	0.1%	97.4%	99.8%	0.21%

Table 2. Comparison of Result Building Modules using posteriors of LSTM

result builder	FPR		recall	F1 score
	in-house	live traffic test set	in-house	all
last-frame	0.1%	0.1%	94.1%	96.3%
last-frame-ignore-last-50	0.1%	0.1%	94.4%	96.5%
median	0.1%	0.3%	97.3%	97.1%
median-ignore-last-50	0.1%	0.8%	97.5%	94.9%
mean	0.1%	0.2%	97.4%	98.2%
mean-ignore-last-50	0.1%	0.3%	97.1%	96.9%

Sum of Residual Harmonics (SRH): Whisper speech is typically characterized by the absence of fundamental frequency (F0) due to a lack of voicing. The SRH feature, originally proposed for robust pitch tracking in noisy conditions by Drugman and Alwan [22], is used as a voicing detector in this work. The SRH feature uses harmonic information in the residual signal and is calculated as:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)] \quad (1)$$

where $E(f)$ is the amplitude spectrum for each Hanning-windowed frame, and for voiced speech presents peaks at the harmonics of F0. The second term in summation, $E((k - \frac{1}{2}) \cdot f))$, helps reduce the relative importance of the maxima of SRH at even harmonics. The value of SRH is sensitive to the initial FFT size, and higher FFT sizes lead to better separation between the values of SRH features in whisper versus normal speech.

High Frequency Energy (HFE): Inspired by the observations in [1] about power spectrum differences in low/high band frequencies between whisper and normal speech, the HFE feature consists of two dimensions. The first dimension reflects the energy ratio between the high frequency band (6875~8000hz) energy and the low frequency band (310~620hz) energy. Whisper generally has less energy in lower frequency bands, thus this ratio can be effective in distinguishing whisper and normal speech. The high and low frequency bands are empirically selected to maximize the separation. The second dimension is the Shannon entropy of the low frequency area. This entropy is calculated by treating the power spectrum as a probability distribution. Whisper tends to have high entropy in the low frequency band.

Auto-Correlation Peak Maximum (ACMAX): The maximum autocorrelation peak within the plausible human F0 range (80~450 hz) is calculated and used as the first dimension for this feature. A value is identified as a peak if it is larger than its 4 neighbors on the left and right. The second and third dimensions of the ACMAX feature consist of the position of the peak and the mean distance between consecutive autocorrelation peaks, respectively.

4.2. Feature and Classifier Evaluation

To evaluate the effectiveness of the features, we trained our LSTM and MLP models with new input vectors consisting of 64 LFBE features plus the 6 engineered features discussed in previous paragraphs. Table 3 shows the evaluation results on in-house and live traffic test sets.

The addition of the engineered features to the existing LFBE features improves both models. For the LSTM model, the engineered features help improve the frame accuracy by a relative 2.6%, leading to more than 99% recall on in-house whisper utterances while reducing rate of false-positives by half. We further observe the LSTM trained only on LFBE features performs comparably to the MLP trained with LFBE and engineered features. This observation suggests that the LSTM model can indeed learn more of the underlying characteristic differences of whisper speech from LFBE features in comparison with the MLP model. While the recall values for the LFBE LSTM are slightly lower than the LFBE + engineered MLP in this case, an improvement in false-positive rate is observed in comparison and the models share similar F1 scores at the same operating point.

For the MLP model, despite the frame accuracy drop in the MLP trained on LFBE + engineered features, the engineered features improve the model performance in terms of

Table 3. Comparison of classifiers trained on LFBE only and LFBE + engineered features.

model	input features	in-house test set			live traffic test set		
		frame acc.	FPR (OP)	recall	frame acc.	FPR	F1 score
MLP	LFBE	77.1%	0.1%	95.9%	94.9%	1.5%	94.9%
LSTM	LFBE	93.5%	0.1%	97.4%	99.7%	0.2%	98.2%
MLP	LFBE + engineered	74.6%	0.1%	98.8%	98.0%	0.6%	98.2%
LSTM	LFBE + engineered	96.0%	0.1%	99.3%	99.9%	0.1%	99.6%

in-house recall, live traffic FPR, and overall F1 score. The drop in MLP frame accuracy could potentially be attributed to a caveat in our data labels where the utterance level tags, i.e. whisper/non-whisper, are propagated to all the frames. This includes both speech frames of interest and non-speech frames such as silence and non-speech noise. In reality, the silence and noise frames need to be labeled separately as they are common and indistinguishable between whisper and non-whisper utterances. The addition of engineered features in the MLP, while not addressing the confusion at the frame-level, seems to be helping to address this issue at the utterance level.

5. CONCLUSIONS

In this work, we proposed using LSTM networks for the task of detecting whisper speech using standard and widely-used LFBE features. We also developed and reviewed a set of features engineered for the task of whisper speech classification and compared the detection ability of the models with and without these engineered features. Our findings show that, with sufficient data, LSTMs can learn the underlying characteristic differences of whisper speech from LFBE features alone, without requiring more sophisticated engineered features. This representational power with standard features makes these LSTMs better candidates for large-scale applications. We show we can further improve the LSTM model performance by utilizing the engineered features in addition to the original LFBE features.

In future work, we plan to experiment with more complex and informed inference modules, including a module using an underlying voice activity detection (VAD) model to filter or weight frames based on their likelihood of containing speech content. We also plan to improve our models' robustness to varied recording conditions and languages by incorporating mixed-condition and mixed-language data into our training and evaluation.

6. REFERENCES

[1] Stanley J Wenndt, Edward J Cupples, and Richard M Floyd, "A study on the classification of whispered and normally phonated speech," in *Seventh International Conference on Spoken Language Processing (ICSLP)*, 2002, pp. 649–652.

[2] Siobodan T Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[3] Johnny B Wilson and James D Mosko, "A comparative analysis of whispered and normally phonated speech using an LPC-10 vocoder," Tech. Rep., Air Force Research Laboratory, Rome,NY, 1985.

[4] Chi Zhang and John HL Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Eighth Annual Conference of the International Speech Communication Association (Interspeech)*, 2007, pp. 2289–2292.

[5] Chi Zhang and John HL Hansen, "An entropy based feature for whisper-island detection within audio streams," in *Ninth Annual Conference of the International Speech Communication Association (Interspeech)*, 2008, pp. 2510–2513.

[6] Chi Zhang and John HL Hansen, "Advancements in whisper-island detection within normally phonated audio streams," in *Tenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2009, pp. 860–863.

[7] Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382.

[8] Roland Maas, Sree Hari Krishnan Parthasarathi, Brian King, Ruitong Huang, and Björn Hoffmeister, "Anchored speech detection," in *Seventeenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 2963–2967.

[9] Won-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee, and Jong-Seok Lee, "Speech/non-speech classification using multiple features for robust endpoint detection," in *IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP), 2000, vol. 3, pp. 1399–1402.

[10] Trausti Kristjansson, Sabine Deligne, and Peder Olsen, "Voicing features for robust speech detection," in *Ninth European Conference on Speech Communication and Technology*, 2005, pp. 369–372.

[11] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park, and Chungyong Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[12] Rathinavelu Chengalvarayan, "Robust energy normalization using speech/nonspeech discriminator for german connected digit recognition," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[13] Adil Benyassine, Eyal Shlomot, H-Y Su, Dominique Massaloux, Claude Lamblin, and J-P Petit, "ITU-T recommendation G. 729 annex b: a silence compression scheme for use with G. 729 optimized for V. 70 digital simultaneous voice and data applications," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[14] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on speech and audio processing*, vol. 10, no. 7, pp. 504–516, 2002.

[15] Samuel Thomas, Sri Harish Mallidi, Thomas Janu, Hynek Hermansky, Nima Mesgarani, Xinhui Zhou, Shihab Shamma, Tim Ng, Bing Zhang, Long Nguyen, et al., "Acoustic and data-driven features for robust speech activity detection," in *Thirteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[16] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association (Interspeech)*, 2014.

[17] R. Maas, A. Rastrow, C. Ma, Lang G., K. Goehner, G. Tiwari, S. Joseph, and B. Hoffmeister, "Combining acoustic embeddings and decoding features for enf-of-utterance detection in real-time far-field speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[18] Yoshua Bengio, Patrice Simard, and Paolo Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[19] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Léon Bottou, "Stochastic gradient learning in neural networks," *Proceedings of Neuro-Nımes*, vol. 91, no. 8, pp. 12, 1991.

[21] Ronald J Williams and Jing Peng, "An efficient gradient-based algorithm for on-line training of recurrent network trajectories," *Neural computation*, vol. 2, no. 4, pp. 490–501, 1990.

[22] Thomas Drugman and Abeer Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics.," in *Twelfth Annual Conference of the International Speech Communication Association (Interspeech)*, 2011, pp. 1973–1976.