Increasing Isolated Word Recognition Performance by Training Models with Reverberant Audio

Fernanda de Souza Ferreira and Tiago Fernandes Tavares

***Abstract*— Isolated Word Recognition (IWR) can be used in different applications, including home automation and car device control. These applications often take place in reverberant environments. Reverberation causes spectral distortion, harming IWR performance. We propose a multi-condition training method that uses both reverberant and non-reverberant audio to improve its generalization capabilities. Reverberant audio is obtained by applying digital sound effects to the training dataset. We used the proposed method to train an existing, baseline IWR system. Results show increased reverberation robustness in various conditions. Therefore, the proposed method poses an important contribution to voice control applications in reverberant environments.**

***Keywords*— Isolated Word Recognition, Reverberation, Multi-Condition Training.**

I. Introduction

Automatic Speech Recognition (ASR) has been employed in various applications, such as car device control [1] and smart homes [2], with the aim of providing comfort, security and accessibility to people. These places are reverberant, that is, sound waves reflect on their boundaries and inner objects. These reflections are summed with the original sound wave producing the reverberation, which decreases speech intelligibility and harms the ASR performance.

Previous researchers have dealt with the improvement of ASR performance in reverberant environments. Some researchers used the dereverberation technique based on spectral subtraction. This technique enhances the speech signal minimizing the acoustic effects before the training stage [3], which it is totally opposite to the technique used in our work. Some researchers have employed multi-condition training technique combined with some methods as CMLLR (Constrained Maximum Likelihood Linear Regression), which helps to reduce the error rate of words [4]. Also, others researchers have employed multi-condition training combined with a similar method as MAP (Maximum a Posteriori), which minimizes unwanted effects caused by background noise [5]. At first, our work just employed the multi-condition training technique using a controlled dataset, seeking similar results.

Multi-condition training uses two types of dataset to train the recognition system: a clean dataset (i.e., data without distortions or effects) and a processed datasets (i.e., data with added reverberation or noise). This shown training process improves generalization capabilities in non-ideal (noisy and reverberant) environments.

School of Electrical and Computer Engineering, University of Campinas (UNICAMP), Campinas-SP, Brazil, E-mails: ferreif1@dca.fee.unicamp.br, tavares@dca.fee.unicamp.br. This work was supported by CNPq and FAPESP.

We propose an analysis of multi-condition training in an environment in which reverberation levels are strictly controlled. For such, we employed an artificial reverberation algorithm and built a dataset with various levels of reverberation. This allowed analyzing the impacts of multi-condition training in several reverberation conditions.

Our IWR system is based on modeling words using Hidden Markov Models (HMMs) [6]. It represents each word from the vocabulary as an isolated HMM. Their hidden states correspond to utterances, while their observations correspond to frame-wise Mel-Frequency Cepstral Coefficients (MFCCs) extracted from audio signals.

The multi-condition training impacts yield results that are slightly lower than those yielded by the best condition scenarios. This means that multi-condition training is able to produce low error rates in diverse conditions, but these error rates are higher than those produced by training solely for the specific test conditions. Nevertheless, multi-condition training reduces the performance difference between the best and worse testing conditions, thus the resulting classifier is more stable than those trained in single conditions.

The remainder of this work is organized as follows. Section II presents the method used in this work. Section III shows the evaluation process and the experimental results. Section IV presents conclusive remarks.

II. Method

The classifier employed in this work uses a distinct HMM to model each word in the vocabulary, as shown in Figure 1. In these models, states roughly represent utterances, hence emissions correspond to output probabilities, which are calculated by observed audio features over time frames given the models. Each model is optimized to recognize a specific word using the Baum-Welch algorithm. In the prediction stage, the classifier yields the label corresponding to the model with the greatest likelihood related to an observation sequence, as calculated using the Viterbi algorithm.

Each word-related HMM used 12 states, which are able to model different utterances found in the vocabulary. Our model uses Gaussian emissions with a diagonal covariance matrix, thus restricting the number of free parameters and consequently reducing the computational complexity. These emissions are related to a generative model for frame-wise MFCCs, which are calculated as follows.

First, the audio signal is divided into 20ms frames, with a 10ms overlap. These frames are multiplied by a Hanning window. Then, we calculate the Discrete Fourier Transform

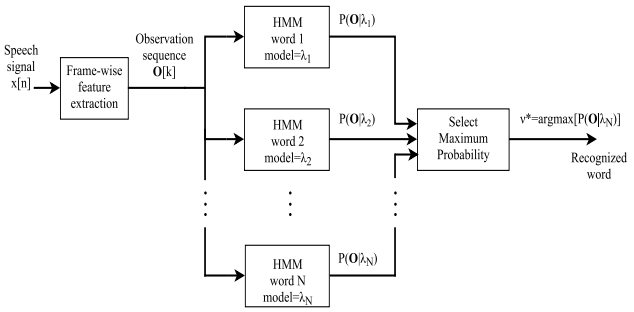

Fig. 1. Block diagram for the IWR system.

(DFT) of each frame. Last, for each absolute value of the DFT, we used 20 triangular bandpass filters to obtain the log energy. These filters are spaced in Mel-frequency, which is related to the common logarithmic frequency. Then the filters outputs have them DCT (Discrete Cosine Transform) calculated, which packs the energies into 30 coefficients for each frame.

The dataset employed in this work aimed at simulating a reverberant environment. For this reason, it comprised recordings of 10 different words spoken in Portuguese. Each word was recorded 120 times by both men and women of different ages, using their own mobile phone microphones. The usage of diverse mobile devices aimed at simulating a non-controlled, real-life application. After recording, each audio track containing a spoken word was subjected to controlled artificial reverberation using the Freeverb algorithm [7]. This process yielded 11 sets with different reverberation levels (from 0 to 100%, with a 10% step). They were labeled R0 to R100, according to their reverberation levels. They were always used symmetrically, that is, if a recording was used for training, then none of its variations would be used for testing. The evaluation method and the results are presented in the next section.

III. Evaluation and Results

This experiment used a train-test evaluation based on a stratified 10-fold cross validation protocol. In each fold, the number of true positives (TP), false positives (FP) and false negatives (FN) for each label (i.e., each word) was calculated. They were used for the estimation of the Recall and Precision, as shown in Expressions 1 and 2.

$$\text{Recall} = \frac{TP}{TP+FN}. \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP}. \quad (2)$$

The F1-score, as shown in Expression 3, was calculated for each fold. The average F1-score across all folds was reported and used as performance measure.

$$F1 = 2\frac{\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}. \quad (3)$$

Our tests aimed at showing the effects of multi-condition training in the word recognition system. Thus, we performed two types of protocols. The first type consisted of using data from one single reverberation level for training. The second type involved using data with two or more reverberation levels for training.

Using preliminary tests, we chose to use R0, R10, R30 and R70 in our experiments. Training with R0 is a baseline because it refers to a non-reverberant environment. R10 is a set that contains just a small amount of reverberation. R30 is a set whose training yielded best results cross all test reverberation levels. Last, R70 was used because training with R30 presents a significant performance decay close to 70% reverberation.

The first test used sets R0 and R10 to highlight the effect of using a small amount of reverberation in multi-condition training. As shown in Figure 2, training with R0 implies in a fast performance decay that accompanies the reverberation increase in the test set. Also, training with R10 implies in a weak performance in a non-reverberant environment. Last, the average F1-score when using multi-condition training (R0+R10) is close to the maximum score obtained when using each set individually, therefore both weaknesses are mitigated.

A similar behavior was observed when using R0 and R30 to build a multi-condition training set. As shown in Figure 3, the performance decays for high reverberation (in training with R0) and no reverberation (in training with R30) are mitigated, at the expense of a lower average F1-score.

Seeking for a better performance, a multi-condition training set with R0 and R70 was used the same way as observed on the previous figures. As shown in Figure 4, when the reverberation levels increase the performance decays gradually in training with R0 and in an opposite way it also happens to R70. As a result of this multi-condition training, it was observed that the performance remains more stable than the previous cases.

In the last case, there are all multi-condition trainings as explained previously and the a new training set which contains R0+R30+R70. In the Figure 5, this new training set is more stable than R0+R10 and R0+R30, but it isn't so stable than R0+R70. However, R0+R30+R70 has a better performance than R0+R70.

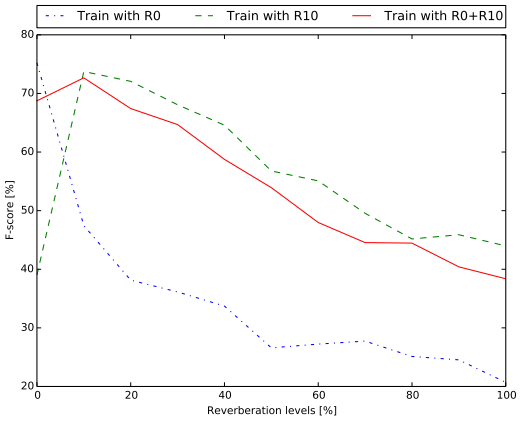

Fig. 2. Classifier performance trained with R0, R10 and R0+R10 datasets.

In both figures 2, 3 and 4, it is possible to observe that multi-condition training yields results with a lower decay when the test set reverberation is increased. This indicates that the training schema yields more stable models than using single-condition training. On the other hand, the best results obtained in multi-condition training are consistently worse than the best

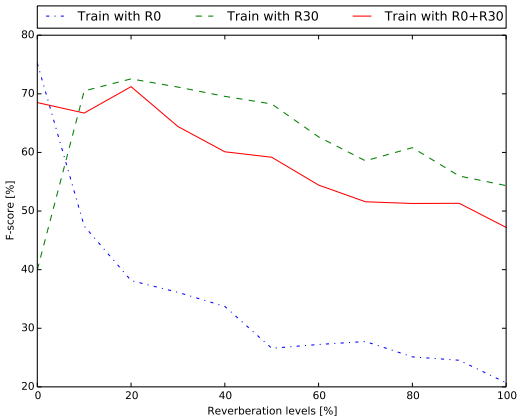

Fig. 3. Classifier performance trained with R0, R30 and R0+R30 dataset.

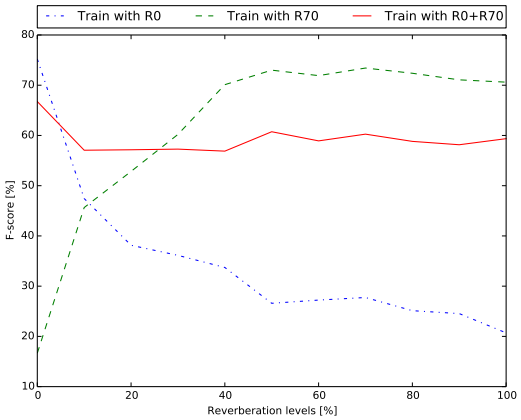

Fig. 4. Classifier performance trained with R0, R70 and R0+R70 dataset.

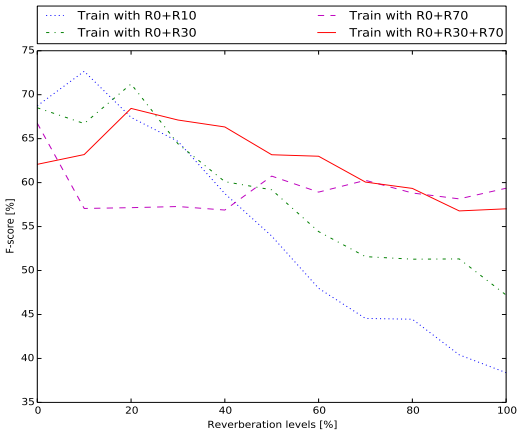

Fig. 5. Classifier performance trained with R0+R10, R0+R30, R0+R70 and R0+R30+R70 dataset.

results obtained in single-condition training. This highlights a trade-off between the model stability and its results.

This trade-off was analyzed using two measures, as shown in Table I. The first one is the difference between the maximum and minimum average F1-Scores for each training set (ΔF). Lower values for this parameter indicate more stable models. The second measure is the sum of all average F1-Scores across different reverberation levels of the test set (ΣF). A higher value for this parameter indicates a higher general performance related to the training set.

Train set	$\Delta F1$	$\Sigma F1$
R0	0.54	3.82
R10	0.34	6.13
R30	0.32	6.84
R70	0.56	6.77
R0+R10	0.34	6.01
R0+R30	0.24	6.45
R0+R70	0.09	6.51
R0+R30+R70	0.11	6.86

TABLE I

DIFFERENCE BETWEEN THE *maxima* AND *minima* OF AVERAGE AND SUM F1-SCORE.

Table I shows that training using two datasets consistently yields models that are more stable (that is, obtained a lower ΔF) that the models obtained by single-condition training. This is an important aspect, because it allows the construction of IWR systems whose behavior is more predictable. However, ΣF in multi-condition training settings do not surpass the best ΣF in the corresponding single-condition settings. This indicates a limitation of the classification algorithm.

Interestingly, Table I shows that training using R0+R70 yields a model that is clearly more stable than others sets($\Delta F = 0.09$). However, training using R0+R30+R70 presents a comparable stability ($\Delta F = 0.11$), but a slightly higher ΣF. Figure 4 shows that good results for training with R0 are obtained in testing setting where training with R70 yields bad results, and vice-versa.

These results show that multi-condition training sets must be chosen among those that generate good results in different conditions. This allows maximizing the resulting system's robustness.

Next section presents conclusive remarks.

IV. CONCLUSIONS

In this work, we evaluated the impact of multi-condition training in the performance of an IWR system in reverberant environment. We used a controlled environment with several reverberation levels that allowed analyzing the IWR in several conditions. Results have shown that this kind of training allows the recognition system's performance to become more stable across different reverberation settings.

Multi-condition training has decreased the system's performance difference between the best and worst case scenarios. However, the performance related to a multi-condition trained system is always slightly worse than the maximum performance related to single-condition training in each condition. This indicates that there is a trade-off between the system's performance in specific environments and the system performance's stability across multiple conditions.

In future work, we plan on adding speech enhancement to a pre-processing stage and increasing the size dataset , therefore potentially improving our classification results.

ACKNOWLEDGEMENTS

Authors would like to thank at CNPq and FAPESP for financial support.

REFERENCES

[1] D. F. Syu, S. W. Syu, S. J. Ruan, Y. C. Huang, and C. K. Yang. Fpga implementation of automatic speech recognition system in a car environment. In *2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*, pages 485–486, 2015.

[2] A. A. Arriany and M. S. Musbah. Applying voice recognition technology for smart home networks. In *2016 International Conference on Engineering MIS (ICEMIS)*, pages 1–6, 2016.

[3] R. Gomez and T. Kawahara. Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1708–1716, 2010.

[4] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas. The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4, 2013.

[5] T. M. Vital and C. A. Ynoguti. Reconhecimento de fala em sistemas veiculares (in portuguese). *Revista Teccen*, 8(2):45–52, 2015.

[6] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[7] Julius O. Smith. *Physical Audio Signal Processing*. https://ccrma.stanford.edu/~jos/pasp/Freeverb.html, 2010. Online book, 2010 edition.