19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

ASSESSING THE REVERBERATION LEVEL IN SPEECH

PACS: 43.72.Ar

Niessen, Maria; Krijnders, Dirkjan; Boers, Joep; Andringa, Tjeerd
Artificial Intelligence, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; m.niessen@ai.rug.nl

ABSTRACT

The performance of automatic speech recognition (ASR) systems is seriously degraded in reverberant environments. We propose a method for assessing the reverberation level in speech that makes it possible to determine in real-time whether a speech signal is reverberant or not. Reverberation causes an increase in the variation of the energy and frequency of harmonics in speech. Speech with a variable pitch is especially affected by reverberation. To capture the effect of reverberation we measured six features on the harmonics of the speech signal, which represent energy and frequency variation in different ways. Speech from the Aurora database was artificially reverberated to demonstrate the validity of these features. Each feature predicted reverberation time for a different subset of the dataset. To test the overall separability of the speech samples using these features, speech from the dataset was automatically classified as being either inside or outside the reverberation radius. Most of the speech was correctly classified, which suggests that a reliable real-time classification algorithm can be developed to select good-quality speech. This algorithm can improve pre-processing methods, such as speech enhancement or voice activity detection, for more robust ASR.

INTRODUCTION

The conditions under which an ASR system is trained will not always match the conditions under which it is used. In real-world environments there may be background noise, other people talking, or effects from room acoustics. The effects from room acoustics are especially disturbing when the speaker does not use a close-talking microphone, and is situated in a reverberant enclosure. All these effects distort the input signal and lead to a degraded performance of the ASR system.

One solution to deal with reverberant environments is to solve the mismatch between the training conditions and the operating conditions. In other words, the ASR system is trained on reverberant speech, and accordingly it will perform better in reverberant operating conditions [1]. However, the effects of room acoustics vary greatly for different environments. Different parameters, such as the size of the room, the material on the floor and walls, and the temperature, influence the acoustic characteristics [2]. Therefore, an ASR system requires training data that match the characteristics of the operating environment. Besides the inconvenience of having to know the operating conditions beforehand, the acquisition of the impulse response describing a room is not straightforward.

Couvreur et al. [3, 4] propose a method where acoustic models are trained on speech under different, simulated reverberant conditions. During operation of the ASR system, the model that matches the operating conditions best is selected. They show an improved performance on simulated reverberated speech compared to an ASR system trained on clean speech. However, the improvement on real data is not as high, because of the discrepancy between real reverberant and simulated reverberant speech.

Other methods try to recover the clean speech from the reverberant signal, for example through a filter that enhances the harmonic structure [5] or pitch-based speech segregation [6]. Although these methods are successful, they are limited to conditions with low levels of reverberation. Furthermore, they rely on the accuracy of the filter that estimates the room impulse response.

We propose a method that classifies speech in a monaural signal as either inside or outside the reverberation radius. Whereas most methods require knowledge of the room characteristics, either for training or filtering, our method allows for blind classification based on properties of the reverberant signal. Reverberation causes an increase in the variation of the energy and frequency of harmonics in speech. Hence, features that capture this variation can be used for real-time classification. We show the validity of six such features.

METHOD

Clean speech was artificially reverberated to enable a controlled experiment. The proposed method required different levels of reverberation for each speech sample, since we examined features that were expected to predict reverberation. Six features are measured in the reverberated speech, and automatic classification is used to test these features.

Reverberation model

A common measure for the level of reverberation is the reverberation time T_{60}. The reverberation time is defined as the time for the sound energy level to decay 60 dB after the excitation has ended. We computed nine different levels of reverberation using the Eyring-Norris equation [7, 8, 2]:

$$T_{60} = \frac{0.161V}{4mV - S\ln(1-\overline{a})} \qquad \text{(Eq. 1)}$$

where V is the room volume in cubic meters; m is a vector with air absorption coefficients for the frequency bands; S is the total surface area; and $\bar{a}$ is the mean wall absorption coefficient. We assumed a fixed room size of 10 by 12 by 3.5 meters and a constant temperature and humidity of respectively 20℃ and 60%. Hence, the mean wall absorption coefficient was the only variable parameter. Values were assigned to this parameter such that the reverberation time intervals were approximately 200 milliseconds.

The reverberation level can also be expressed by the reverberation radius or distance [2]. The reverberation radius is the distance of the speaker or microphone to the sound source for which the energy contribution of the direct sound and the echoes are equal. A more reverberant environment concurs with a smaller reverberation radius. Naturally, the reverberation radius is strongly correlated to the reverberation time. We also compute the reverberation radius so the sound samples can be labeled as either clean or reverberant. In this paper, we regard clean speech as speech inside the reverberation radius.

The parameter values used in the Eyring-Norris equation were used as input to the shoebox model, which simulates an impulse response in a rectangular room, a shoebox. The shoebox model is an implementation of the image source method of Allen and Berkley [9]. The speaker and the listener or microphone are modelled as two points in space. Apart from the direct sound, specular reflections are computed using mirrored image sources. An impulse response is obtained for every image source. The final impulse response describing the room is computed by combining all individual impulse responses, which are received at different delay times. This impulse response is convolved with the speech signal, resulting in reverberant speech. The speech is processed using a cochleogram, a logarithmic time-frequency representation based on a transmission-line model of the human cochlea [10]. Figure 1 depicts a cochleogram of a clean speech sample on the left, and a cochleogram of a reverberant speech sample on the right.

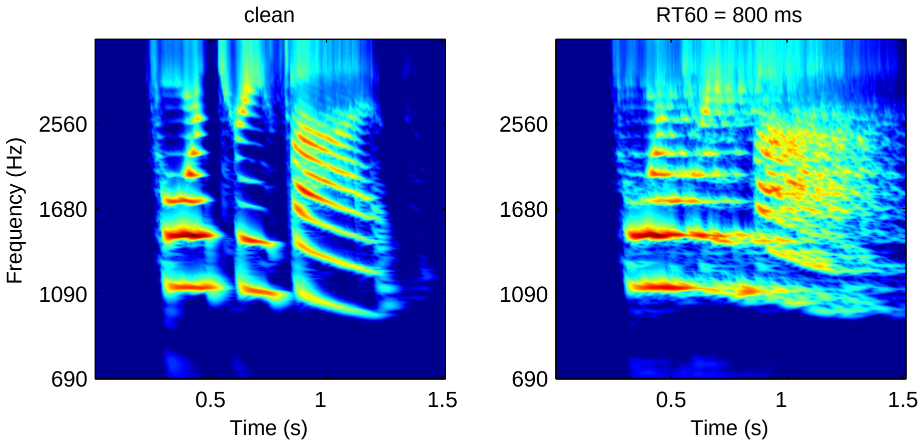

Figure 1. - Cochleograms of '435' by a female speaker, clean and reverberant

Measuring reverberation effects

We expected that the effect of reverberation on speech can be measured directly in the speech signal. Since we want to develop a real-time speech classification system, the features used for the classification must have no parameters that require knowledge of the room characteristics. One prominent effect of reverberation on speech is the attenuated salience, that is, the attenuated stability in both the frequency and the energy, of the harmonics [11]. First, voiced speech was located based on the selection of harmonic complexes — a superposition of co-occurring harmonics — in the cochleogram. Next, the fluctuation in energy and frequency of the first five harmonics of the harmonic complex was measured, because these harmonics are most robust in the cochleogram.

The energy variation is measured through the number of peaks on the harmonic (1) and the energy variation of the harmonic compared to its smoothed version (2). Both values are expected to increase at higher reverberation levels. In addition, the energy contributions of echoes cause less distinct harmonics. This effect is captured by calculating the energy slope of the harmonics (3), and the width of the harmonic compared to an ideal sinusoid (4). The energy slope will be less steep in reverberant speech, and the harmonic will cover a broader frequency range. Reverberation effects can be found in the time-frequency plane as well. The short-time development of the harmonic is distorted by echoes, causing a less smooth harmonic track. Therefore, the track variation was measured compared to its smoothed version (5), and to an approximation of a clean harmonic track (6). These six features are summarized in figure 2.

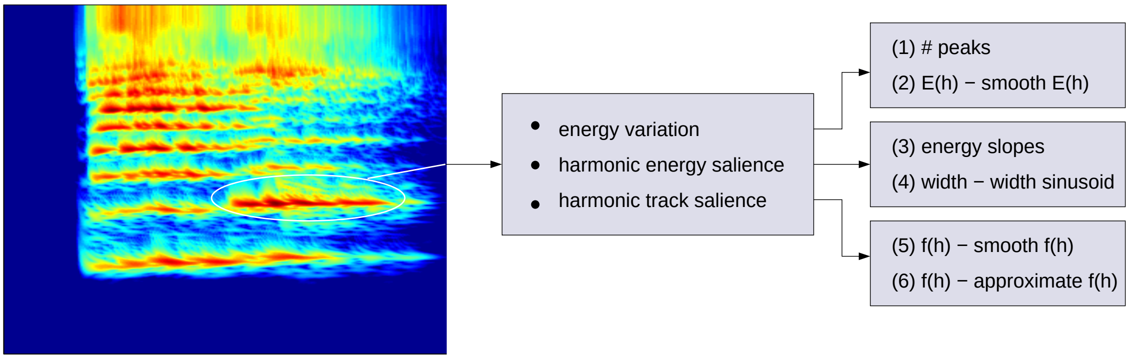

Figure 2. - Features that indicate reverberation level on an harmonic (h)

Figure 3 shows an example of the feature values measured on a single harmonic. The horizontal axis corresponds to the reverberation time in all plots, but the vertical axis is different for each plot, depending on what the feature represents. Note that segments correspond to filter frequency bands. Two of the plots depict two features, because they have similar values on the same dimension. The individual points are single measurements, and the lines are linear fits. For this harmonic, the linear relation between the features values and the reverberation time is significant for all features except the energy variation, *E(h) - smooth E(h)*. In general, not each feature of each sound sample will predict reverberation. Particular characteristics of the sound, such as stationarity or the formant frequencies, change the measurability of the features. However, it is likely that in each sound sample at least one feature can be measured that predicts reverberation, since the features are measured on five harmonics.

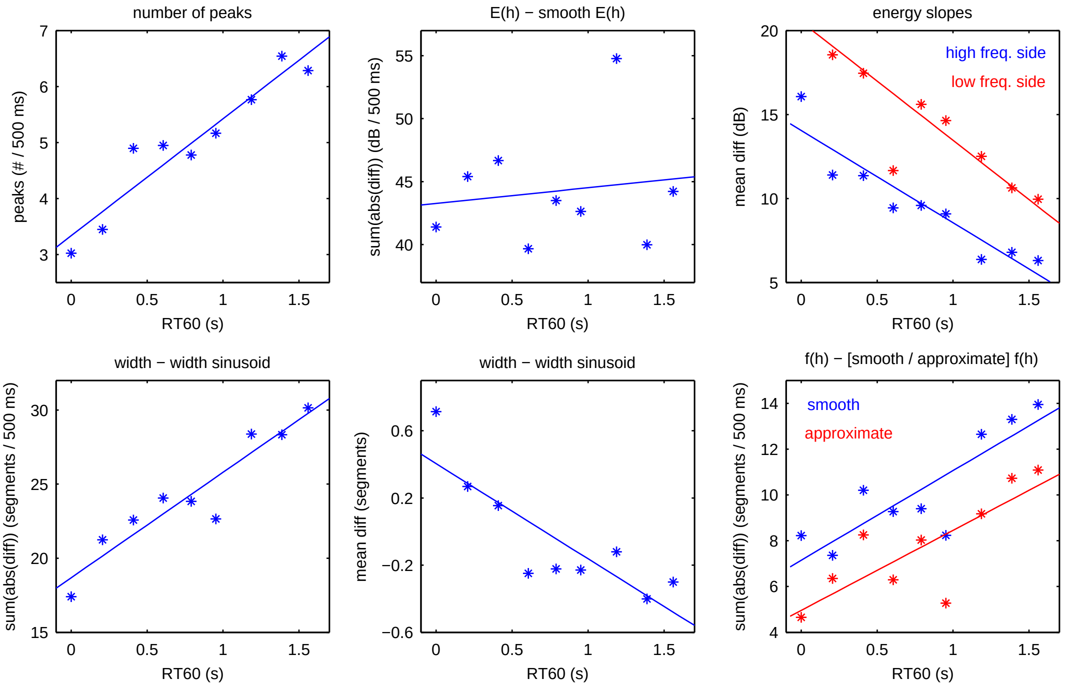

Figure 3. - Feature values of the first harmonic (h) of a short phrase

Speech dataset

Part of the Aurora database [12] was used to validate the six features. Artificial reverberation was added to 685 randomly selected clean sound samples, spoken by 214 different speakers, both male and female. The reverberation was computed at nine levels, equivalent to linear spread reverberation times between 0 and 1600 ms. As we expected, most of the 685 sound samples showed a significant correlation of at least one feature with the reverberation time. Only 6% did not show a relation for any of the features. However, the predictive strength of the features for individual sound samples is no direct indication for the general separability of the speech samples as either clean or reverberant. To test the separability, global thresholds need to be determined in a training set and used to classify a test set.

Since the 685 speech samples were reverberated at nine levels, a total of 6165 samples could be used for classification. After the dismissal of unsound samples, that is, samples in which we could not measure one or more features, 5189 samples were left. All samples within the reverberation radius, the two lowest levels, were labeled as clean, and all samples outside the reverberation radius, the other seven levels, were labeled as reverberant. The data was split in a part for training (33%) and a part for testing (66%). In addition, continuous read speech of six speakers was recorded using a close-talking microphone. This data was split into samples of similar length to the Aurora database, and resampled to an equal sample frequency. The speech samples were artificially reverberated in the same way as the other data. Again, part of the data that was unfit was removed, and 2377 samples were left. These samples served as an extra test set, which can show the robustness of the features.

Classification method

Numerous methods exist to test the classification accuracy of features. We used a support vector machine (SVM) [13], since it is known to be less prone to problems of overfitting than some other methods [14]. Of course, we want our classifier to be as general as possible, because it should work in a variety of environments. In training, an optimal separating hyperplane, or threshold boundary, is determined. The support vectors are the speech samples that are closest to the hyperplane, and hence are most difficult to classify. The mapping of the data to an higher-dimensional space is dependent on the type of kernel, that is, the mapping function, which can be defined by the user, or selected from one of the standards. For our data we use a standard linear kernel. The number of support vectors is an indication of the complexity of the classification. During the testing phase, the speech samples are mapped onto these support vectors. Since the test samples are labeled as well, the classification can be compared to the labels, resulting in a performance measure.

RESULTS

The speech samples from the Aurora database were split randomly into a training set of 1744 samples and a test set of 3445 samples. The six features, measured in eight different ways (see figure 3), were computed on the first five harmonics, resulting in 40-dimensional data. The skewedness of the data — 22% of the speech samples was clean and 78% was reverberant, because the reverberation radius corresponds to a relatively low reverberation time of just over 200 ms — is accounted for by using prior probabilities to weight the class error contributions. The SVM was trained on the training samples, resulting in a classifier with 363 support vectors. The rest of the speech samples of the Aurora database was tested on the trained classifier. The performance, or accuracy, of the classifier was 92%. The additional speech samples of our own recordings were also tested on the classifier, with a performance of 87%, only a few percent less.

Different classification methods could be chosen for this problem, or different settings for the SVM. For example, if the size of the training set is increased, the performance on the other Aurora speech samples increases, but the performance on the extra test set decreases. We are not interested in optimizing the classifier on a particular dataset, but in the separability of any reverberant speech using the features. Hence, the classification with an SVM using a linear kernel gives an indicative performance result.

CONCLUSIONS

We discussed whether the reverberation level can be assessed in speech, without any knowledge of the room characteristics. Six features that were expected to reflect reverberation were measured in speech samples under different, simulated reverberant conditions. We showed that these six features can be used to classify speech as either clean or reverberant. Although the reverberation assessment was reduced to a two-class problem in this study, the predictive strength of the features indicate that a quantitative assessment is possible as well. The possibility of quantitatively determining the reverberation level will be investigated in a follow-up study. Furthermore, we will look into the effects of reverberation on higher harmonics, since the attenuated harmonicity will probably be quite evident there.

So far, we used artificial reverberation to validate the features, which allows controlled measurements at different levels of reverberation. The next step is to test the performance on real reverberant data. When the general validity of the features is shown, we can improve the performance of ASR systems by selecting good quality speech. The comparable results of the two different test sets on the same classifier already suggest that the features are quite robust. Besides the possibility to select speech directly for ASR systems, the features can select speech for pre-processing methods as well, such as the speech enhancement methods discussed in the introduction [3, 4, 5, 6]. The thresholds can be changed depending on the possibilities of the next processing step. Once the applicable thresholds are determined, the classifier can operate real-time, without any knowledge of the operating environment, and without any extra demands on the ASR system.

As mentioned in the introduction, ASR systems should be able to operate in reverberant and noisy environments. In the future, we will try to explicate the physical origin of reverberation effects. We already showed the predictive strength of the features we studied. If we can link specific effects of reverberation to specific features, we can separate the effects of reverberation and background noise, making it feasible to eliminate both.

Acknowledgments

This work is supported by SenterNovem, Dutch Companion project grant nr: IS053013. We are grateful to Koen Haak, Marius Bulacu, Axel Brink and Tijn van der Zandt for their help with SVM.

References: [1] M. Matassoni, M. Omologo, D. Giulini: Hands-free speech recognition using a filtered clean corpus and incremental HMM adaptation. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing **3** (2000) 1407-1410

[2] H. KuttRuff: Room Acoustics. Applied Science Publishers, London (1979)

[3] L. Couvreur, C. Couvreur, C. Ris: A corpus-based approach for robust ASR in reverberant environments. Proceedings of the Sixth International Conference on Spoken Language **1** (2000) 397-400

[4] L. Couvreur, C. Couvreur: Blind model selection for automatic speech recognition in reverberant environments. Journal of VLSI Signal Processing **36** (2004) 189-203

[5] T. Nakatani, M. Miyoshi, K. Kinoshita: Blind dereverberation of monaural speech signals based on harmonic structure. Systems and Computers in Japan **37** (2006) 1-12

[6] N. Roman, D. Wang: Pitch-based monaural segregation of reverberant speech. The Journal of the Acoustical Society of America **120** (2006) 458-469

[7] C. F. Eyring: Reverberation time in "dead" rooms. The Journal of the Acoustical Society of America **1** (1930) 168

[8] R. F. Norris: An instrumental method of reverberation measurement. The Journal of the Acoustical Society of America **1** (1929) 32

[9] J. B. Allen, D. A. Berkley: Image method for efficiently simulating small-room acoustics. The Journal of the Acoustical Society of America **65** (1979) 943-950

[10] H. Duifhuis, H. Hoogstraten, S. Netten, R. Diependaal, W. Bialek: Modeling the cochlear partition with coupled Van Der Pol oscillators. In J. W. Wilson, D. T. Kemp (Eds.) Cochlear Mechanisms: Structure, Function and Models (1985) 395-404

[11] C. J. Darwin, R. W. Hukin: Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention. The Journal of the Acoustical Society of America **108** (2000) 335-342

[12] Aurora Project Database. http://www.elda.org/

[13] PRTools. http://www.prtools.org/

[14] R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification. John Wiley & Sons, New York (2001)