ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Descriptive Vocabulary Development for Degraded Speech

Dushyant Sharma[1], *Gaston Hilkhuysen*[2], *Patrick A. Naylor*[1], *Nikolay D. Gaubitch*[1], *Mark Huckvale*[2] *and Mike Brookes*[1]

Centre for Law Enforcement Audio Research,
[1] Electrical and Electronic Engineering, Imperial College London, UK
email:dushyant.sharma02@imperial.ac.uk

[2] Speech, Hearing & Phonetic Sciences, University College London, UK
email:g.hilkhuysen@ucl.ac.uk.com

10.21437/Interspeech.2012-423

Abstract

This paper presents the development of a compact vocabulary for describing the audible characteristics of degraded speech. An experiment was conducted with 51 English-speaking subjects who were tasked with assigning one of a list of given text descriptors to 220 degradation conditions. Exploratory data analysis using hierarchical clustering resulted in a compact vocabulary of 10 classes, which was further validated by a bootstrap cluster analysis.

Index Terms: speech degraded vocabulary, speech quality, correspondence analysis

1. Introduction

Speech signals for telecommunications and surveillance are often degraded by the acoustic environment in which they are captured and by non-ideal operation of the signal acquisition and transmission systems.In audio processing for surveillance and law enforcement, there is a requirement from audio analysts and practitioners for a concise vocabulary to describe the perceived characteristics of a degraded speech signal. This paper presents the clustering of a large vocabulary of text descriptors into classes with similar perceptual characteristics such that words in the same class can be considered as near synonyms. This work aims at facilitating consistent and repeatable description of degraded speech such as would allow an audio transcriber to identify and communicate the audible characteristics of any degraded speech encountered. Further analysis of the signal properties of the audio associated with each class could be used to select or suggest the best enhancement regime for the corresponding degraded signals. This work relates closely to studies on multidimensional speech quality. The diagnostic acceptability measure (DAM) [1] was proposed in 1977 as a multidimensional speech quality framework, where the quality of the signal is characterized by 10 perceptual scales, 6 for describing the perceptual qualities of the foreground and 4 for the background. The DAM rating form presented in [2] characterizes the speech quality on a scale with 16 attributes. Similar studies were presented in [3] and [4]. The scope of this study emphasizes degradations commonly found in surveillance and law enforcement audio and is limited to native English-speaking subjects. Analysis of the data using hierarchical clustering under the framework of correspondence analysis results in a concise vocabulary for describing the perceptual effects of the degradations and a bootstrapping validation is used to identify a robust clustering solution.

2. Methodology

2.1. Initial vocabulary and preparation for the TAXIT experiment

A first pilot study was conducted on 6 expert subjects in which they were tasked to label degraded speech examples using one of 46 labels extracted from the DAM [1] and Matilla [3] studies. The subjects were asked to provide any additional labels that they would have liked to use for each of the 220 test conditions. The results from the pilot study suggested the addition of the “noisy” and “natural” labels to the entire vocabulary. The resulting 48 labels are presented using the experiment interface as shown in Fig. 1. Following this pilot study, the Taxonomy Labeling Experiment (TAXIT) was conducted using the enhanced label set and is described in the following parts of this paper.

2.2. Database

Audio stimuli consisting of 220 sentences spoken by a male speaker were employed in the experiments [5]. A total of 55 base degraded speech conditions, denoted C01 to C55, were established, as described below:

- Clean (C01 - 02): undegraded speech.
- Brick-wall filtering (C03 - C18) : low-pass, band-pass and high-pass filters with 50 Hz transition bands and stop-band attenuation of 60 dB applied.
- Coloration (C19 - C32) : shelf filters with low and high cut and boost as well as 2 types of spectral tilt.

- Additive noise (C33 - C41): car, babble and hum noise were added to the speech at signal-to-noise ratios (SNR) of -5, 0 and 5 dB.
- Reverberation (C42 - C44): Room impulse responses from the MARDY database [6] with reverberation time (T60) of 0.60 seconds and direct-to-reverberation ratios of 18.8, 19.2 and 19.9.
- Envelope fluctuations (C45 - C46): two random fluctuations in the speech envelope were applied.
- Clicks and dropouts (C47 - C52): temporal erasures were applied to randomly selected speech segments in the signal.
- Peak clipping (C53 - C55): symmetric hard clipping was applied with thresholds of -20, -25 and -30 dBFS.

Each of the base conditions were processed by GSM 6.10 (13 kbps), GSM transcoding (GSM 6.10 followed by G.711 followed by GSM 6.10) and MP3 (16 kbps) CODECs and used in addition to linear PCM versions. All audio was sampled at 8 kHz and peak normalized to -10 dBFS.

2.3. Subjects

A total of 51 naive, native English speaking subjects between the age of 20 and 50 years were recruited for the experiments. All subjects reported normal hearing and were paid for their participation. The listening tests were conducted in a sound-proof booth, with stimuli presented via Sennhieser HD 650 headphones driven by an RME Fireface 800 digital-to-analogue converter. Subjects received instructions on the task.

2.4. Procedure

The subjects were presented with 48 text descriptors arranged on a graphical user interface according to similarity. The task was to identify the best text descriptor for "perceived quality of the audio". The presentation order of the stimuli was randomized between subjects and the average time for completing the task was 45 minutes, including a 5 minute break half way through the task. The presentation gain for the stimuli was set by the subjects at the beginning of the experiment to a "comfortable level" and all subsequent stimuli were presented to the subject at this level.

3. Analysis

The subjective data was analyzed using correspondence analysis, which is an exploratory analysis tool for categorical response data [7]. Since the objective of the experiment was to cluster the responses into classes of text descriptors representing similar perceptual characterization, a hierarchical clustering technique was applied to the data to discover the classes of vocabulary [8].

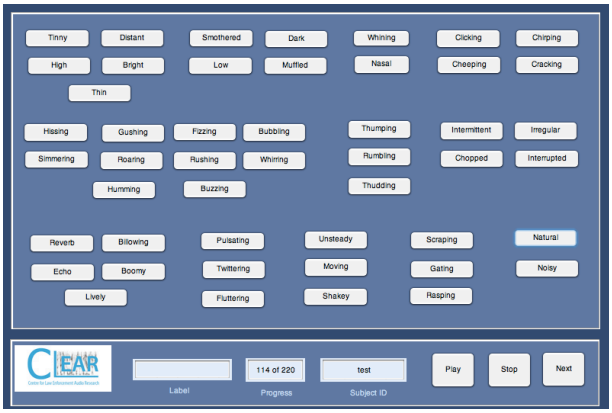

Figure 1: TAXIT labeling experiment interface.

3.1. Correspondence analysis

Here we present the correspondence analysis of the TAXIT results. The results are characterized by an $I \times J$ matrix $\mathbb{X}$ representing the vocabulary as the rows ($I = 48$) and the degradation conditions as the columns ($J = 220$). Each element of this matrix represents the frequency of selection of the text descriptor $(i = 1, 2, \ldots, I)$ for condition $(j = 1, 2, ..., J)$. The row and column masses of $\mathbb{X}$ are calculated as follows:

$$y_{ij} = \frac{x_{ij}}{\sum_i \sum_j x_{ij}} \qquad r_i = \sum_j y_{ij} \qquad c_j = \sum_i y_{ij}. \tag{1}$$

The vector $\boldsymbol{c}^{\mathrm{T}}$ represents the average row profile (or centroid) of the data matrix,and $(.)^{\mathrm{T}}$ indicates matrix transposition. The profile a_{ij} of the j^{th} element of row i is defined as

$$a_{ij} = \frac{x_{ij}}{r_i}. \tag{2}$$

In correspondence analysis, distances are measured by the chi-squared statistic (χ^2). The χ^2 distance between row i and the centroid of $\mathbb{X}$ is calculated as

$$d_{i,c} = \sqrt{\sum_{j=1}^{J} \frac{(a_{ij} - c_j)^2}{c_j}}. \tag{3}$$

The inertia of the results matrix is a measure of the variability of the row profile a_i relative to the centroid, calculated as:

$$\Theta^2 = \sum_{j=1}^{J} r_i d_{i,c}^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} (a_{ij} - c_j)^2 / c_j. \tag{4}$$

These quantities will be used in the merging procedure of the correspondence analysis as will be described in the following section.

3.2. Clustering vocabulary

A hierarchical clustering method is applied based on the Ward clustering algorithm [8]. The rows of the matrix $\mathbb{X}$ are successively merged, beginning with the full matrix (all rows separate) and continuing until only one row remains (all rows merged). When two rows are merged, the change in inertia of the merged matrix may be decomposed into the between-groups inertia (total inertia of the merged table) and within-groups inertia (reduction in inertia when two rows are merged). The criteria is to maximize the between-groups inertia and minimize the within-groups inertia [8]. This is equivalent to minimizing the following measure:

$$\lambda_{i,i'} = \frac{r_i r_{i'}}{r_i + r_{i'}} d^2_{i,i'} \tag{5}$$

where r_i and $r_{i'}$ are the row masses corresponding to the rows being merged (i.e. rows i and $i^{'}$) and $d^2_{i,i'}$ is the χ^2 distance between the rows. The hierarchical clustering algorithm partitions the results matrix into a maximum of I clusters and minimum of 1. There are a number of methods to identify the number of classes or clusters that are present in the data. A common technique is to find the "knee" in the scree plot. Figure 2 shows a plot of the reduction in inertia against the number of classes for the clustering of the vocabulary.

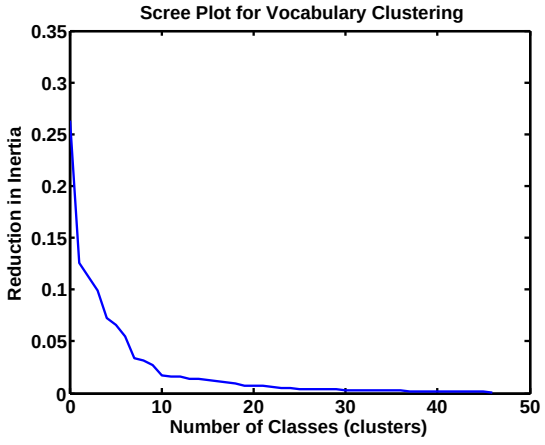

Figure 2: Scree plot for the TAXIT experiment.

3.3. Cluster Validity

Although a knee based technique is useful in identifying the number of classes, it is not guaranteed that this represents the most stable clustering solution as it may be a result of the particular sampling of the subjects from the population. A number of alternate techniques exist that exploit the data available to validate the stability of the clustering solution to variations in the data. The bootstrapping techniques construct subsamples of the data (considering the data as the population) without replacement and apply a figure of merit (FOM) to establish the reliability of the clustering solution for different number of clusters. The particular method we employed is applicable to any clustering algorithm and is an example of bootstrapping cluster validity [9]. Let the number of subjects be denoted by N and let V be the number of clusters then, τ_{ij} is an $N \times N$ connectivity matrix defined as follows:

$$\tau_{i,j} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The method proceeds by creating m ensembles of τ_{ij} using $f \times N$ subsets, where f represents a dilution factor (set to $2/3$, resulting in 34 subjects per subset). The plot of the average FOM (1000 resamples) against the number of clusters is presented in Fig. 3.

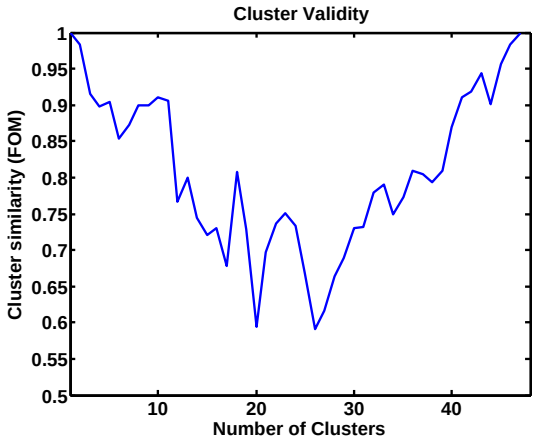

Figure 3: Average figure of merit for the hierarchical clustering algorithm applied to the TAXIT data.

3.4. Principal descriptors

Apart from determining and validating the number of clusters in the data it is also desirable to determine an appropriate descriptor for each cluster based on some criterion. The first criterion we choose is the minimum χ^2 distance from each word in a cluster to its centroid, referred here as the Type A criterion. This measure identifies the principal descriptor as the label that is closest to the centroid of the cluster. Another criterion is the minimum of the ratio of the within class to the between class χ^2 distance for each word in a cluster. This metric is referred as the Type B criterion and identifies the principal descriptor as the label that is most distinct. The resulting principal descriptors are shown in Table 1.

4. Results

The analysis of the vocabulary assignment by 51 naive listeners results in the clustering of the responses into a 10 class vocabulary as shown in Fig. 2. The knee in the scree plot suggests that 10 classes in the data are sufficient to account for 95% of the inertia in the data. A cluster validation also confirms that the 10 class clustering solution is stable as shown in Fig. 3 (indicated by the local maximum of the FOM at 10 clusters).

In addition to the clustering of the vocabulary, results for two criteria for describing the principal descriptors for each class are shown in Table 1. The type A principal descriptors for the filtering and coloration degradations are shown in Table 2. This labeling allows us to describe the perceptual effects of the degradations, for example, low pass filtered speech is perceived as "Muffled" when transmitted through a PCM channel, and changes to the "Pulsating" class of descriptors when a GSM CODEC is present.

Cluster Number	Principal Descriptor	
	Type A	Type B
1	Cracking	Cracking
2	Buzzing	Buzzing
3	Natural	Natural
4	Noisy	Noisy
5	Humming	Humming
6	Muffled	Muffled
7	Interrupted	Interrupted
8	Rushing	Hissing
9	Distant	Tinny
10	Pulsating	Cheeping

Table 1: Principal descriptors for the 10 classes.

Degradation	PCM	GSM	MP3
Clean	Natural	Pulsating	Pulsating
Low pass	Muffled	Pulsating	Pulsating
High pass	Tinny	Pulsating	Pulsating
Band pass	Distant	Pulsating	Pulsating
Low shelf	Muffled	Muffled	Pulsating
High shelf	Muffled	Muffled	Pulsating

Table 2: Type A principal descriptors for the filtering and coloration conditions (C01 to C32).

5. Conclusions

This paper presented a vocabulary development experiment for labeling degraded speech signals, conducted on 51 naive native English subjects using 220 degradation conditions relevant to the surveillance and law enforcement audio processing field. Exploratory data analysis correspondence analysis led to a clustering of the 48 label vocabulary into 10 classes. This result was further validated by performing a bootstrapping analysis using a figure of merit on 1000 resamples of the data. The result showed that the 10 class clustering solution was stable to sample fluctuations. Additionally, results for two methods of determining a label from the vocabulary to serve as the principal descriptor for each cluster were presented. The concise label vocabulary provides for the identification and communication of the audible aspects of degraded speech on a 10 label vocabulary.

6. References

[1] W. Voiers, "Diagnostic acceptability measure for speech communication systems," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1977, pp. 204–207.

[2] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Prentice Hall, Jan. 1988.

[3] V.-V. Mattila, "Semantic analysis of speech quality in mobile communications: descriptive language development and mapping to acceptability," *Food Quality and Preference*, vol. 14, pp. 441–453, November 2003.

[4] M. Waltermann and A. R. andS.Moller, "Perceptual dimensions of wideband-transmitted speech," in *2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems*, Germany, September 2006, pp. 103–108.

[5] M. W. Smith and A. Faulkner, "Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing," *J. Acoust. Soc. Am.*, vol. 120, pp. 4019–4030, 2006.

[6] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, Sep. 2006.

[7] H. Abdi and L. J. Williams, *Encyclopedia of Research Design*, N. Salkind, Ed. Thousand Oaks (CA): Sage, 2010.

[8] M. Greenacre, *Correspondence Analyis in Practice*, 2nd ed. Chapman & Hall/CRC, 2007.

[9] E. Levine and E. Domany, "Resampling method for unsupervised estimation of cluster validity," *Neural Computation*, vol. 13, pp. 2573–2593, 2001.