Noisy_TIMIT

The Continuous Speech of DARPA TIMIT in Various Noise Levels

Azhar Abdulaziz, Veton Këpuska
Florida Institute of Technology, ECE Department, Melbourne, FL 32901

1.Introduction

The Noisy_TIMIT corpus is a new version of the well-known TIMIT corpus. This work was a part of the research that has been conducted on automatic speech recognition in noisy environment.

It is designed to simulate the audio utterances for different additive noise levels. Only the audio has been modified, so that the original arrangement of the TIMIT corpus is still as described by the TIMIT documents.

TIMIT is a US English continuous speech corpus that is phonetically balanced. It also utilizes a large number of most common accents in the United States. For more information about the corpus design, the following section is *COPIED* from the original TIMIT Readme.doc file (Garofolo 1993).

2.TIMIT Corpus Design

2.1 Corpus Speaker Distribution

TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. Table 1 shows the number of speakers for the 8 dialect regions, broken down by sex. The percentages are given in parentheses. A speaker's dialect region is the geographical area of the U.S where they lived during their childhood years.

The geographical areas correspond with recognized dialect regions in U.S. (Language Files, Ohio State University Linguistics Dept., 1982), with the exception of the Western region (dr7) in which dialect boundaries are not known with any confidence and dialect region 8 where the speakers moved around a lot during their childhood.

Dialect Region(dr)	#Male	#Female	Total
1	31 (63%)	18 (27%)	49 (8%)
2	71 (70%)	31 (30%)	102 (16%)
3	79 (67%)	23 (23%)	102 (16%)
4	69 (69%)	31 (31%)	100 (16%)
5	62 (63%)	36 (37%)	98 (16%)
6	30 (65%)	16 (35%)	46 (7%)
7	74 (74%)	26 (26%)	100 (16%)
8	22 (67%)	11 (33%)	33 (5%)
Total	438 (70%)	192 (30%)	630 (100%)

The dialect regions are:

dr1: New England

dr2: Northern

dr3: North Midland

dr4: South Midland

dr5: Southern

dr6: New York City

dr7: Western

dr8: Army Brat (moved around)

2.2 Corpus Text Material

The text material in the TIMIT prompts (found in the file "prompts.doc") consists of 2 dialect "shibboleth" sentences designed at SRI, 450 phonetically-compact sentences designed at MIT, and 1890 phonetically-diverse sentences selected at TI. The dialect sentences (the SA sentences) were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either difficult or of particular interest. Each speaker read 5 of these sentences (the SX sentences) and each text was spoken by 7 different speakers. The phonetically-diverse sentences (the SI sentences) were selected from existing text sources - the Brown Corpus (Kuchera and Francis, 1967) and the Playwrights Dialog (Hultzen, et al., 1964) - so as to add diversity in sentence types and phonetic contexts. The selection criteria maximized the variety of allophonic contexts found in the texts.

Each speaker read 3 of these sentences, with each sentence being read only by a single speaker. Table 2 below summarizes the speech material in TIMIT.

Table 2: TIMIT speech material

Sentence Type	#Sentences	#Speakers	Total	#Sentences/Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342		6300	10

2.3 Suggested Training/Test Subdivision

The speech material has been subdivided into portions for training and testing. The criteria for the subdivision is described in the file "testset.doc".

Core Test Set:

The test data has a core portion containing 24 speakers, 2 male and 1 female from each dialect region. The core test speakers are shown in Table 3. Each speaker read a different set of SX sentences. Thus the core test material contains 192 sentences, 5 SX and 3 SI for each speaker, each having a distinct text prompt.

Table 3: The core test set of 24 speakers

Dialect	Male	Female
1	DAB0, WBT0	ELC0
2	TAS1, WEW0	PAS0
3	JMP0, LNT0	PKT0
4	LLL0, TLS0	JLM0
5	BPM0, KLT0	NLP0
6	CMJ0, JDH0	MGD0
7	GRT0, NJM0	DHC0
8	JLN0, PAM0	MLD0

Complete Test Set:

A more extensive test set was obtained by including the sentences from all speakers that read any of the SX texts included in the core test set. In doing so, no sentence text appears in both the training and test sets. This complete test set contains a total of 168 speakers and 1344 utterances, accounting for about 27% of the total speech material. The resulting dialect distribution of the 168 speaker test set is given in Table 4. The complete test material contains 624 distinct texts.

Table 4: Dialect distribution for complete test set

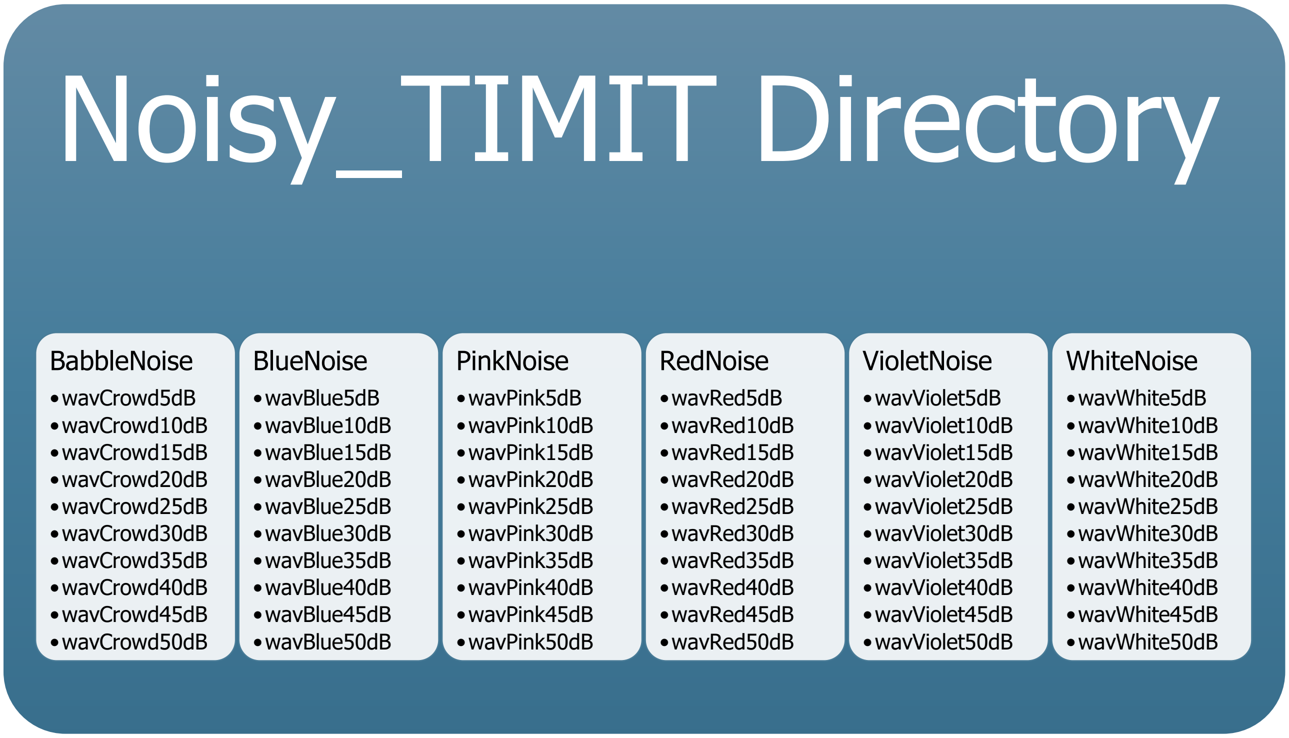
Dialect	#Male	#Female	Total
1	7	4	11
2	18	8	26
3	23	3	26
4	16	16	32
5	17	11	28
6	8	3	11
7	15	8	23
8	8	3	11
Total	112	56	168

3.Noise Types and Ranges

The additive noise is designed to be white, pink, blue, red, violet and babble noise. Noise types have been artificially generated using MATLAB programing environment except for the babble noise. The technical description of the noise generation and addition is discussed in section 5 below. The noise level changes is based on the Signal-to-Noise Ratio (SNR) provided in decibel (dB). The noise level is varying in 5 dB steps and ranges from 5 to 50 dB.

4.Noisy_TIMIT Directory Structure

The following figure shows the main structure for Noisy_TIMIT directory.



Noisy_TIMIT Directory

BabbleNoise

- wavCrowd5dB
- wavCrowd10dB
- wavCrowd15dB
- wavCrowd20dB
- wavCrowd25dB
- wavCrowd30dB
- wavCrowd35dB
- wavCrowd40dB
- wavCrowd45dB
- wavCrowd50dB

BlueNoise

- wavBlue5dB
- wavBlue10dB
- wavBlue15dB
- wavBlue20dB
- wavBlue25dB
- wavBlue30dB
- wavBlue35dB
- wavBlue40dB
- wavBlue45dB
- wavBlue50dB

PinkNoise

- wavPink5dB
- wavPink10dB
- wavPink15dB
- wavPink20dB
- wavPink25dB
- wavPink30dB
- wavPink35dB
- wavPink40dB
- wavPink45dB
- wavPink50dB

RedNoise

- wavRed5dB
- wavRed10dB
- wavRed15dB
- wavRed20dB
- wavRed25dB
- wavRed30dB
- wavRed35dB
- wavRed40dB
- wavRed45dB
- wavRed50dB

VioletNoise

- wavViolet5dB
- wavViolet10dB
- wavViolet15dB
- wavViolet20dB
- wavViolet25dB
- wavViolet30dB
- wavViolet35dB
- wavViolet40dB
- wavViolet45dB
- wavViolet50dB

WhiteNoise

- wavWhite5dB
- wavWhite10dB
- wavWhite15dB
- wavWhite20dB
- wavWhite25dB
- wavWhite30dB
- wavWhite35dB
- wavWhite40dB
- wavWhite45dB
- wavWhite50dB

Figure 1: Noisy_TIMIT basic directory hierarchy

In each subdirectory in figure 1, the same structure for directory region (dr) and sentence type file names is used in this corpus. However, this structure will be repeated for each noise level and type. There are six types of noise and for each noise category there are 10 different noise levels. The phonetic description files are not provided. Only the prompts are provided in separate files.

The hierarchical file arrangement then is identical to the original TIMIT corpus. Starting from "<Noisy_TIMIT>/<Noise_Type>/wav<Type>xdB>/", the speech flac files are organized according to the following hierarchy (Garofolo 1993):

<USAGE>/<DIALECT>/<SEX><SPEAKER_ID>/<SENTENCE_ID>.<FILE_TYPE>

where,

USAGE :== train | test

DIALECT :== dr1 | dr2 | dr3 | dr4 | dr5 | dr6 | dr7 | dr8 (see Table 1 for dialect code description)

SEX :== m | f

SPEAKER_ID :== <INITIALS><DIGIT>

where,

INITIALS :== speaker initials, 3 letters

DIGIT :== number 0-9 to differentiate speakers with identical initials

SENTENCE_ID :== <TEXT_TYPE><SENTENCE_NUMBER>

where,

TEXT_TYPE :== sa | si | sx (see Section 2.2 for sentence text type description)
SENTENCE_NUMBER :== 1 ... 2342

FILE_TYPE :== flac

Example: If you have the following file: Noisy_TIMIT/RedNoise/wavRed15dB/train/dr1/fcjf0/sa1.flac, it means that:

Audio with additive red noise, training set, dialect region 1, female speaker, speaker-ID "cjf0", sentence text "sa1", speech waveform file compressed using flac.

Then the text associated with this audio is found in the transcription file /NoisyTIMIT_Documentation/timit_train.transcription, in a line that ends with (SA1).

Note that, this line is repeated in this file because its associated audio is repeated multiple of times in the training set. The test and train transcription files are the text representation for the audio directories and arranged in the order of occurrence.

5.The Noise Generation and Addition

The white, pink, blue, red and violet noise types are generated artificially using MATLAB. The function `randn()`, which generates normally distributed random numbers, is used to generate the white noise. For other colored noise types the following relationships are used (RANE CORPORATION n.d.) to transform the white noise to the following:

Pink noise $\mathrm{S(f)} \propto 1/\mathrm{f}$

Blue noise $\mathrm{S(f)} \propto \mathrm{f}$

Red noise $\mathrm{S(f)} \propto 1/\mathrm{f}^2$

Violet noise $\mathrm{S(f)} \propto \mathrm{f}^2$

Where $S(f)$ is the Power Spectral Density (PSD) of the normally distributed random signal.

After calculating the signal power of the TIMIT clean audio, the resultant noise signals powers are scaled to obtain certain signal-to-noise ratios (SNR). After scaling the noise power, the noisy audio is formed by adding the clean audio to the noise as follows:

$x(t) = s(t) + n(t)$ ………………………. (1)

where $s(t)$ is the original TIMIT audio signal, $n(t)$ is the white or colored noise signal.

For the babble noise, a random segment from a recorded babble speech noise is selected and its power is scaled relative to the power of the original TIMIT audio signal. Equation 1 above is used to get the final noisy audio signal $x(t)$.

References

Garofolo, J. S., L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett. 1993. *"Darpa timit acoustic–phonetic continous speech corpus cd-rom documentation. .* CD, DARPA, USA: NIST.

RANE CORPORATION. n.d. *RANE Audio Products for Profissionals.* Accessed 2015. http://www.rane.com/par-n.html#noise_color.