STC-TIMIT: Generation of a Single-channel Telephone Corpus

Nicolás Morales[1], Javier Tejedor[2], Javier Garrido[2], José Colás[2], Doroteo T. Toledano[3]
[1]Nuance Communications GmbH. Kackerstrasse, 10, Aachen, Germany.
[2]HCTLab-Escuela Politécnica Superior, Universidad Autónoma de Madrid.
[3]ATVSLab-Escuela Politécnica Superior, Universidad Autónoma de Madrid.
Francisco Tomás y Valiente, 11. Madrid, Spain.
E-mail: nicolas.morales@nuance.com, {javier.tejedor, javier.garrido, jose.colas, doroteo.torre}@uam.es

Abstract

This paper describes a new speech corpus, STC-TIMIT, and discusses the process of design, development and its distribution through LDC. The STC-TIMIT corpus is derived from the widely used TIMIT corpus by sending it through a real and single telephone channel. TIMIT is phonetically balanced, covers the dialectal diversity in continental USA and has been extensively used as a benchmark for speech recognition algorithms, especially in early stages of development. The experimental usability of TIMIT has been increased eventually with the creation of derived corpora, passing the original data through different channels. One such example is the well-known NTIMIT corpus, where the original files in TIMIT are re-recorded after being sent through different telephone calls, resulting in a corpus that characterizes telephone channels in a wide sense. In STC-TIMIT, we followed a similar procedure, but the whole corpus was transmitted in a single telephone call with the goal of obtaining data from a real and yet highly stable telephone channel across the whole corpus. Files in STC-TIMIT are aligned to those of TIMIT with a theoretical precision of 0.125 ms, making TIMIT labels valid for the new corpus. The experimental section presents several results on speech recognition accuracy.

1. Introduction

The DARPA-TIMIT Acoustic-Phonetic Continuous Speech Corpus, commonly known as TIMIT (Fisher *et al.*, 1986) is a benchmark corpus in the field of Automatic Speech Recognition (ASR). It consists of phonetically labeled short sentences of English read speech under clean conditions, and 630 speakers covering the dialectal variability in the USA. The practical value of this corpus has been expanded over time with the creation of derived corpora where the original speech files are re-recorded under different conditions or distortions. In the LDC catalogue (LDC) the following related corpora are currently found: NTIMIT (original files sent through the Public Switched Telephone Network (PSTN) in different calls), CTIMIT (re-recording of files sent through cellular telephones), HTIMIT (different transducers are employed) and FFMTIMIT (the secondary release of the original TIMIT recorded with a free-field microphone). These derived corpora complement the original TIMIT corpus allowing for extensive experimental research and comparison of performance under a variety of conditions. On the other hand, the new corpora benefit from TIMIT by sharing the same labeling information.

This paper presents a new derived corpus, STC-TIMIT, where the original files are sent trough a real and single telephone channel. This is the key procedure that differentiates STC-TIMIT from NTIMIT and makes it unique. In NTIMIT (Jankowski *et al.*, 1990) each original speech file is sent through a different telephone call and as a result, each of them is affected by a different channel distortion. This is in spite of the standards of the ITU for PSTN networks (ITU-T, 2001), which nevertheless are flexible. Therefore, calls with different origins and destinations follow different channel distortions (even if the origin and destination are the same, the conditions of the line may vary significantly in different calls, especially in long-distance ones that involve multiple carrier companies). On the contrary, the newly created STC-TIMIT, while corresponding to a real telephone channel, presents a much more stable and well-defined channel response across files. While NTIMIT is adequate for testing speech algorithms in a varied set of telephone channels, STC-TIMIT presented here is superior for the study of algorithms in which it is assumed that a single channel affects speech, or when sufficient amounts of data from different distortions are required.

Two other important differences exist between NTIMIT and STC-TIMIT; firstly, in NTIMIT an artificial mouth and a handset were used for transmission of the original files to the telephone channel, while in STC-TIMIT the original files are directly transmitted by means of a telephone switchboard. Secondly, alignment with the original TIMIT database is more precise in STC-TIMIT than in NTIMIT (Sections 2 and 3). As we show in the experimental section, misalignment may have an impact in performance for particular applications.

The rest of this chapter is organized as follows. In Sections 2 and 3 we describe the methodology and characteristics of NTIMIT and STC-TIMIT, respectively. In Section 4 we show ASR accuracy experiments using TIMIT, NTIMIT and STC-TIMIT, respectively and in Section 5 we discuss the distribution of the new corpus STC-TIMIT. Conclusions are presented in Section 6.

2. Methodology and Characteristics of NTIMIT

The PSTN is the sum of the world's public circuit-switched telephone networks, ruled by the

At the time of writing, the first author was with HCTLab and ATVSLab at Universidad Autónoma de Madrid, Spain.

standards of the ITU-T. In the United States the network is organized in Local Access and Transport Areas (LATAs). When a call is made between 2 numbers in the same LATA the call is solely handled by the local telephone company. On the contrary, a call connecting different LATAs, is first handled by the originating local company, then by a long-distance carrier and finally by the receiving local company, in what is termed a long-distance call.

In NTIMIT calls originate from the NYNEX Science and Technology Laboratories in White Plains, New York. Half of the utterances were transmitted within the local LATA and the other half reached one of 10 selected LATAs. Thus, a wide range of channel conditions are represented in the database (in average each combination of originating and receiving points has a representation of 77 seconds).

A known issue with NTIMIT is a small misalignment between the original files (from TIMIT) and their re-recorded equivalents. As explained in (Jankowski *et al.*, 1990), for the purpose of alignment with TIMIT, sets of clicks were added at the start and end of utterances. However, it is reported that this method produced misalignments in approximately 10% of utterances. Misaligned utterances had to be retransmitted, and four passes were necessary for correct alignment of the complete corpus. Nevertheless, there were some minor issues with the resulting corpus, such as incomplete utterances (at least 3 in the training set and 6 in the test set[1]) and a small misalignment. We measured the average misalignment per utterance by maximizing the cross-correlation between TIMIT and NTIMIT files, resulting 3.7 ± 0.9 ms (59.5 ± 15.0 samples). Thus, if the signal is parameterized using, for example, a window size of 25 ms, the average misalignment represents 15% of the window size. In Section 4 we show that for some applications this misalignment may have a significant impact on system performance.

3. Methodology of STC-TIMIT

STC-TIMIT was recorded by passing the original files of the TIMIT corpus through a real telephone channel (this does not include a telephone handset, as the signal was directly transmitted to the telephone channel using a telephone switchboard). The database is organized in the same manner as the original TIMIT corpus: 4620 files belonging to the training partition and 1680 belonging to the test partition. Files are recorded using 8 kHz sampling frequency and μ-Law encoding. Four sets of calibration tones were added at the start of every quarter of the whole database, consisting of:

- A 1 kHz tone of duration 4 seconds.
- A sweep tone from 10 Hz to 4000 Hz of duration 4 seconds.

Similar calibration tones are also given in the NTIMIT corpus and allow characterizing the telephone channel conditions in which each utterance is transmitted.

The following shows the process of preparation, execution and post-processing of STC-TIMIT.

[1] This problem was reported by PhD. B. Pellom.

3.1 Dialogic Switchboard and Telephone Channel

Speech utterances were sent through the telephone network and recorded by means of a Dialogic D/41JCT-LS switchboard (D41JCT-LS). One of the 4 integrated telephone lines was used as the caller end (sending speech data to the network) and another one as the receiving end (recording data). The process was handled using a voice platform designed in our group (Tomico *et al.*, 2003), capable of executing a variety of automatic telephone services. The telephone interface server module was developed using Intel's Dialogic System Release 5.1.1 software for Windows. The two lines were inside the building of the Escuela Politécnica Superior (Universidad Autónoma de Madrid) and the call was therefore handled locally.

In NTIMIT the signal was passed to the telephone line using a handset and an artificial mouth placed in an acoustically isolated room (Jankowski *et al.*, 1990). On the contrary, in STC-TIMIT we use the switchboard to generate directly the telephone signal at the calling end and therefore it is not affected by convolutional distortions caused by the telephone microphone.

Sampling frequency for recording of STC-TIMIT is 8 kHz and speech files are stored in μ-Law format.

3.2 Data Preparation, Recording and Calibration Tones

A single audio file was created by concatenation of each file in the TIMIT corpus. However, given the limitation of total recording time in the Dialogic switchboard of a maximum of 6000 seconds, the process was divided in 4 fragments of approximately 4800 seconds (total size of TIMIT is around 19380 seconds). Each of the four subsets was preceded by two calibration tones: a fixed 1000 Hz tone and a linearly varying tone from 10 to 4000 Hz, both with duration of 4 seconds (this is similar to the calibration tones existing in NTIMIT for each LATA). All 4 fragments were sent through the line in a single call, *i.e.*, no hang-up was made between recording of each of them. Thus, in principle, the channel should remain stable throughout the duration of the recording of the whole database. Nevertheless, temporal variations may exist and these may be assessed by means of the calibration tones.

In Figure 1 we show four spectrograms for sweeping-tone calibration files: 2 from NTIMIT and 2 from STC-TIMIT. From the spectrograms on the top, it is clear that STC-TIMIT presents a very stable band-pass filtering distortion with attenuation near the borders. This is a consequence of all files (and calibration tones) being sent through a unique local telephone channel. Such stable spectrograms mean that all source files will be affected by a very similar distortion. On the contrary, the two spectrograms from NTIMIT (bottom) present a more complicated distortion shape due to the combination of multiple elements (artificial mouth, telephone handset and telephone channel). In addition, they are more dissimilar to each other, because each of them corresponds to a different call. This poses a more

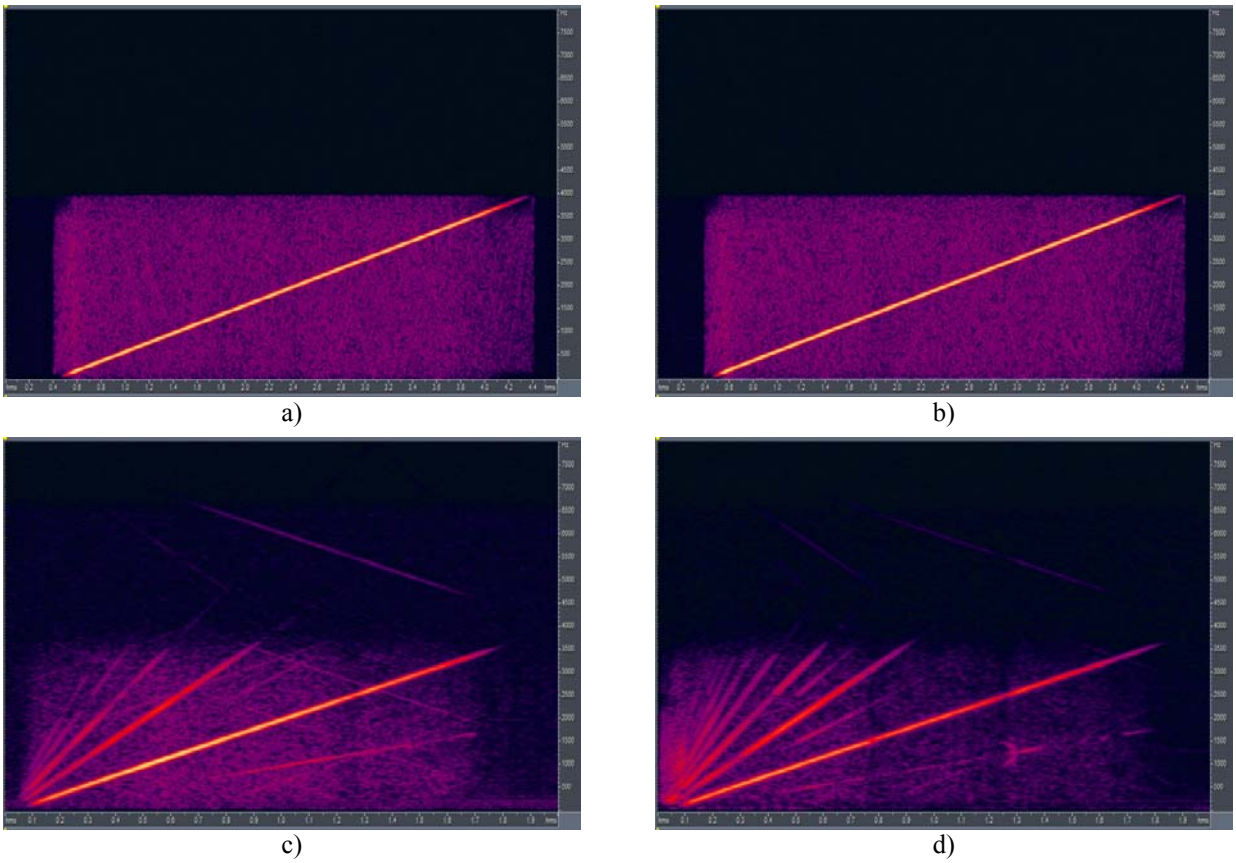

Figure 1: Spectrograms of randomly chosen calibration tones for STC-TIMIT (*a* and *b*) and NTIMIT (*c* and *d*). STC-TIMIT presents a very stable band-pass filtering distortion with attenuation near the borders, as a result of passing the source calibration tones through a unique local telephone channel. On the contrary, NTIMIT presents a more complicated distortion shape due to the combination of multiple elements (artificial mouth, telephone handset and telephone channel). NTIMIT calibration tones are more dissimilar, because each of them corresponds to a different call, from a common origin but with a large variety of destinations.

challenging problem when the goal is to model the bandwidth limitation distortion, or to compensate it, for example for speech recognition tasks. For the above reasons, NTIMIT and STC-TIMIT share the common characteristic of representing telephone versions of TIMIT, but result in significantly different levels of distortion of the original data and both represent significant complements to TIMIT. STC-TIMIT can be interpreted as a mild and stable distortion of TIMIT due to the bandwidth limitation and small noise; a distortion introduced by a local phone call, while NTIMIT includes variable amounts of distortions, also including the effect of the telephone handset. The four calibration tones used for this analysis were chosen randomly and are not particular in any respect. Therefore the preceding analysis is valid in general for any pairs of utterances from NTIMIT and STC-TIMIT.

3.3 Post-processing Alignment with TIMIT

Utterances in the new corpus were originally obtained by splicing the single-call-recorded file according to the file sizes in TIMIT. However, we observed that this method generated misalignments due to the slight difference between play output and recording sampling rates (the difference is approximately 1 sample every 16 seconds, or equivalently, 1 sample every 128000 recorded samples; this is irrelevant for a particular utterance, but makes it impossible to align the whole corpus with this method). Therefore, a more sophisticated approach was employed; in the release version of STC-TIMIT alignment was made individually for each utterance by maximizing the cross-correlation function computed between the original TIMIT file and its corresponding recorded file (Matlab function *xcorr*, from the Signal Processing Toolbox was used), which in theory would yield a precision of ± 1 sample (0.125 ms) per utterance. It should be noted that the previous statement is based on the assumption that the best possible alignment between two utterances is that maximizing their cross-correlation (precisely the criterion used for post-processing).

3.4 Summary of Distortions Present in STC-TIMIT

There are two main sources of distortion:

- Digital/Analog - Analog/Digital conversions: The original digital signal is converted to analogical format by the D/A converter in the Dialogic switchboard, introducing a distortion not present in the original TIMIT

Corpus	No CMN		With CMN	
	% Corr.	% Acc.	% Corr.	% Acc.
TIMIT	75.40	71.18	75.71	71.61
STC-TIMIT	69.10	61.80	69.52	62.15
NTIMIT	62.45	53.76	64.07	55.73

Table 1: Phone-based recognition rates for TIMIT, STC-TIMIT and NTIMIT, respectively, using matched training and testing conditions.

files. Similarly, at the receiving end an additional distortion is introduced by the A/D converter.

- Channel effects: These depend on the specific characteristics of the telephone channel used. Given the complexity of a telephone network these effects are difficult to estimate and the best way to characterize them is by means of calibration tones or comparison of original and re-recorded files.

4. Experimental Results on Automatic Speech Recognition

The principal intent in the creation of STC-TIMIT was for experimentation on robustness algorithms against bandwidth limitation for Automatic Speech Recognition (ASR).

The major cause of degradation of ASR in STC-TIMIT is the effect of the telephone channel, similar to band-pass filtering. Several studies on machine and human recognition show that most of the important signal information in the English language is contained in the spectral region that goes from 300 Hz to 3400 Hz (Warren *et al.*, 1995; Allen, 1994). Nevertheless, studies on the spectral distribution of particular groups of phonemes show that some important differentiating information may be contained in the upper frequencies, especially for fricatives (Junqua & Haton, 1996; Huang *et al.*, 2001; ITU, 1993).

Another possible source of degradation is the difference in the distorting channel for different speech files that may cause mismatches between training and testing conditions and would require more robust acoustic models, for example incrementing the number of Gaussian mixtures. As previously discussed, we expect this mismatch to be small in STC-TIMIT because all speech files were transmitted in a single telephone call and the upper spectrograms in Figure 1 show that the distortion is stable in time. In fact, we assume the degradation in STC-TIMIT due to this mismatch to be negligible. On the contrary, STC-TIMIT and NTIMIT having similar average cut-off frequencies and SNRs, we can have a qualitative idea of the impact of small variations of the telephone channel by comparing accuracies of matched systems for these two corpora.

In Table 1 we compare ASR recognition using matched training and testing conditions and identical experimental setups for TIMIT, NTIMIT and STC-TIMIT. A phonetic recognition engine based on Hidden Markov Models (HMM) is trained for each of them, and a common and a phone bigram language model is employed. Fifty-one acoustic models (3 emitting states with 15 Gaussian mixtures each) are trained including short-pause and starting and ending silences. The front-end uses pre-emphasis filtering (α=0.97) and 25 ms Hamming windows with a 10 ms window shift. Thirteen MFCC coefficients including C0 and their respective first and second order derivatives (39 total features) are computed from a filter-bank of 26 triangular filters uniformly distributed in the Mel-Frequency scale along the region 0-8 kHz. HMM models are trained using the whole training partition of TIMIT. Table 1 shows phone-based percent correct and accuracy for each corpus, with and without Cepstral Mean Normalization (CMN).

Corpus	Compensation mode	% Corr.	% Acc.
Aligned STC-TIMIT	Multivariate-32	62.53	56.78
	Multivariate-256	64.67	58.79
Misaligned STC-TIMIT	Multivariate-32	61.45	55.78
	Multivariate-256	63.91	58.07

Table 2: Speech recognition performance for multivariate compensation trained with two versions of STC-TIMIT: the first one is perfectly aligned with TIMIT and the other has the same misalignment as NTIMIT.

The difference between rows 1 and 2 (TIMIT and STC-TIMIT) shows approximately the impact of the telephone channel in a simple ASR system. This degradation is mostly due to loss of information in the signal and is therefore difficult to compensate. The difference between rows 2 and 3 (STC-TIMIT and NTIMIT) is due to both, a slightly more distorting channel (in spectrograms *c* and *d* in Figure 1 we see non-linear distortions producing harmonics), and the variability of this channel. The degradation due to variability may, in principle, be alleviated by means of robustness approaches.

As it was stated in Section 3.3 an important feature of STC-TIMIT is its highly accurate alignment to TIMIT. A direct application of this alignment is that phonetic labels from TIMIT are also valid for STC-TIMIT. Additionally, particular applications that require high precision alignment may also benefit from this feature. For example, we were interested in stereo-based compensation techniques of band-limited speech for ASR. The goal in this type of task is to perform speech recognition of band-limited speech such as in STC-TIMIT or NTIMIT, using acoustic models trained with full-bandwidth data, as that from TIMIT. In such cases there is a mismatch between training and testing conditions and a compensation of the input signal is required. Stereo-based approaches make use of simultaneously recorded speech signals under band-limited and full-bandwidth conditions in order to learn a compensation and apply this for incoming test utterances (Morales *et al.*, 2007a; Morales 2007). It is reasonable then that learning of the transformation between the full-bandwidth and limited bandwidth spaces will be more effective if the two corpora are aligned precisely.

In Section 2 we reported an average misalignment in

NTIMIT of 3.7 ms (or 15% of a window of length 25 ms). In Table 2 we evaluate what would be the impact of such misalignment in the STC-TIMIT corpus. Performance measures are given for multivariate stereo-based feature compensation (Morales *et al.*, 2007b) using 32 and 256 classes and two mentioned versions of STC-TIMIT: aligned and misaligned. Results in Table 2 show that there is a significant degradation in ASR performance for a system where multivariate feature compensation is made using compensation classes trained with misaligned data. The previous experiment is only one example of the utility of a precise alignment for corpora derived from TIMIT, which may nevertheless find many other applications. The details of the experimental setup described are out of the scope of this paper and can be consulted in (Morales, 2007).

5. Distribution through LDC

STC-TIMIT has been approved for publication in the LDC catalogue (LDC) and its release is scheduled for April 2008. The release will contain all files in the original TIMIT corpus re-recorded in the new conditions and the calibration tones.

6. Conclusions

We have presented the process of design and creation of STC-TIMIT. The new corpus complements its parent corpus TIMIT, and other derived corpora by introducing a new distortion: a unique and real telephone channel applied to all speech files. Alignment of the new corpus to TIMIT is optimal and assures coherence of phonetic labeling, as well as optimal performance on situations requiring precise alignment such as stereo-based algorithms.

Experimental results presented show that the coherence of the distortion across all files in STC-TIMIT allows for better acoustic modeling than in TIMIT where each file was sent through a different call (therefore different channel). We also showed the advantage of a precise alignment between corpora.

7. References

Allen, J.B. (1994). How Do Humans Process and Recognize Speech? *IEEE Transactions on Speech and Audio Processing*, vol. 2 (4), pp. 567--577.

D41JCT-LS switchboard datasheet. http://download.intel.com/design/telecom/prodbref/6925.pdf

Fisher, W.M., Doddington, R., Goudie-Marshall, K.M. (1986). The DARPA speech recognition research database: specifications and status. *Proceedings of the DARPA workshop on Speech Recognition*, pp. 93--99.

Huang, X., Acero, A., Hon, H.W. (2001). Spoken language processing: a guide to theory, algorithm and system development, pp. 36--51. Prentice Hall.

ITU (1993). Paired comparison test of wideband and narrowband telephony. Technical Report COM 12-9-E, ITU. March, 1993.

ITU-T (2001). Transmission performance characteristics of pulse code modulation channels. Recommendation G. 712 (11/01).

Jankowski, C., Kalyanswamy, A., Basson, S., Spitz, J. (1990). NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. *Proceedings of ICASSP'90*, vol. 1, pp. 109--112.

Junqua, J.C., Haton, J.P. (1996). Robustness in Automatic Speech Recognition, pp. 10--20. Kluwer Academic Publishers.

LDC catalogue. http://www.ldc.upenn.edu/

Morales, N., Toledano, D.T., Hansen, J.H.L., Colás, J. (2007a). Blind feature compensation for time-variant band-limited speech recognition. *IEEE Signal Processing Letters*, vol. 14 (1), pp. 70--73.

Morales, N., Toledano, D.T., Hansen, J.H.L., Garrido, J. (2007b). Multivariate cepstral feature compensation on band limited data for robust speech recognition. *Proceedings NODALIDA'07*, pp. 144--151.

Morales, N. (2007). Robust Speech recognition under band-limited channels and other channel distortions. PhD. Dissertation, Computer Science Department. Universidad Autónoma de Madrid, Spain.

Tomico, V., Morales, N., Campos, E., Tejedor, J., Bolaños, D., Jiménez, S., Garrido, J., Colás, J. (2003). Vocal platform for telephone voice portals and internet based interfaces. Proceedings m-ICTE 2003, pp. 1889--189

Warren, R.M., Riener, K.R., Bashford, J.A., Brubaker, B.S. (1995). Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. *Perception & Psychophysics*, vol. 57 (2), pp. 175--182.