NTIMIT: A Phonetically Balanced, Continuous Speech, Telephone Bandwidth Speech Database

Charles Jankowski, Ashok Kalyanswamy, Sara Basson, and Judith Spitz

NYNEX Science and Technology Center
White Plains, New York

ABSTRACT

The creation of a continuous speech, multi-speaker, telephone bandwidth speech database is described. The NTIMIT (Network TIMIT) database was collected by transmitting the TIMIT database over the telephone network. Additional advantages of the NTIMIT database include a carefully selected diversity of speech dialects and extensive breadth and depth of phonetic coverage. NTIMIT is orthographically and phonetically labelled identically to the TIMIT data. Possible uses for NTIMIT include acoustic analysis of telephone bandwidth speech, development of telephone bandwidth speech recognition algorithms, and retraining current wideband algorithms for telephone speech. Speech transmission was achieved by creating a "loopback" telephone path to a large number of central offices. The central offices were geographically distributed to simulate different telephone network conditions. Half of the TIMIT database was sent over "local" telephone paths, while half was transmitted over "long distance" conditions. Transmission involved the use of a commercial device to simulate the acoustic characteristics between a human's mouth and a telephone handset. All recordings were done in an acoustically isolated room. Calibration signals were transmitted to each central office in order to readily evaluate such network characteristics as attenuation, frequency response, and harmonic distortion.

1. Introduction

Critical to the performance of any speech recognition system is a large amount of speech data well matched to the speech recognition application at hand. For a limited or isolated application, the optimal choice is to collect data in the actual environment of the application. The long-term goal of speech recognition, however, is to build speaker-independent systems that could function in a multitude of environments and conditions. To build these systems, and to study general characteristics of speech, it is useful to have data that is reflects the general nature of speech and is not designed for any particular application.

The TIMIT database is an example of this type of speech database. Useful characteristics of the database include many speakers, continuous speech, carefully controlled selection of speech dialects from around the Continental United States, and carefully designed breadth and depth of phonetic coverage. TIMIT is not useful for telephone bandwidth speech analysis or algorithm development, though, since the speech was recorded in careful acoustically controlled conditions with a high quality wideband microphone.

It is well known that the telephone network creates both linear and nonlinear distortions to the speech signal during transmission through the network. Such effects include low and high-pass filtering, spectral distortion, additive noise, and nonlinear harmonic distortion. Here the "telephone network" includes such equipment as telephone microphones, handsets, and phone sets themselves, as well as the actual loop or trunk conditions.

For telephone bandwidth speech recognition research, therefore, it is desirable to have a database with the advantages of the TIMIT database, coupled with the effects introduced by the telephone network. These advantages can be achieved by transmitting the TIMIT database over the telephone network. Equipment exists to simulate the necessary components of the telephone transmission stage, including the acoustic characteristics of a human mouth and the acoustic path between the mouth and the telephone. This equipment allows simulating the TIMIT database being recorded originally over the telephone, which is the desired result.

This paper describes the creation of the Network TIMIT, or NTIMIT database, which is the result of transmitting the TIMIT database over the telephone network. In the first section, a brief description of the TIMIT database is given, including characteristics useful for speech analysis and recognition. Next, the hardware and software required for the transmission of the database is described. The geographic distribution of the TIMIT utterances is described, followed by information regarding calibration signals used to readily determine the effect of the transmission setup and the distortions introduced by the telephone network.

2. TIMIT Database

A full description of the TIMIT database can be found in [1]. The database consists of 630 speakers, 438 male and 138 female. The speakers are categorized into one of eight "dialect regions" that approximately map to speech dialects in American English.

Each speaker spoke ten utterances, each of which was one sentence. Two of the ten sentences were "dialect sentences" that were spoken by every speaker. These sentences were intended to gauge the effect of dialect on various characteristics of speech. Five sentences were "MIT" sentences, iteratively designed by hand with the goal of providing a rich variety of phonetic segments and phonetic contexts. The remaining three sentences were "TI" sentences. These sentences were designed by taking a large corpus of written text (the Brown Corpus) and augmenting it with a small number of selections from a corpus of "spoken" sentences. Then, an automatic design process was instituted which pruned the resulting corpus on the basis of richness of phonetic coverage,

CH2847-2/90/0000-0109 $1.00 © 1990 IEEE

until the desired number of sentences remained. There were 450 MIT sentences, each spoken by seven speakers, and 1890 TI sentences, each spoken by one speaker. From the dialect sentences to the MIT sentences to the TI sentences, the phonetic depth of the sub-database decreases, while the phonetic breadth increases.

Two versions of the TIMIT database exist; one with a close-talking Sennheiser microphone, and another with a far-field pressure microphone. Only the close-talking microphone version of the database was used for creating NTIMIT.

Once recorded, the 6300 TIMIT utterances were orthographically and phonetically transcribed. A description of this process appears in [2].

3. Hardware and Network Configuration

Figure 1 shows the hardware configuration used in the transmission setup. Both digital audio devices process speech at a 16 KHz sampling rate 16 bit PCM. The anti-aliasing filter for recording is a 9th order elliptic filter with a cutoff frequency of 6.4 KHz. The TIMIT utterance is transmitted in an acoustically isolated room through an "artificial mouth" designed to approximate the acoustic characteristics of the human mouth. The artificial mouth is mounted in a device designed to calibrate telephone handsets. This "telephone test frame" thus carefully controls such physical parameters as angle of telephone handset relative to the artificial mouth and the distance and angle of the telephone microphone from the output of the artificial mouth. The effect of this device is to approximate the acoustic coupling between the human mouth and the telephone handset. Results presented later in the paper show the effects of the use of the artificial mouth and the telephone handset test frame.

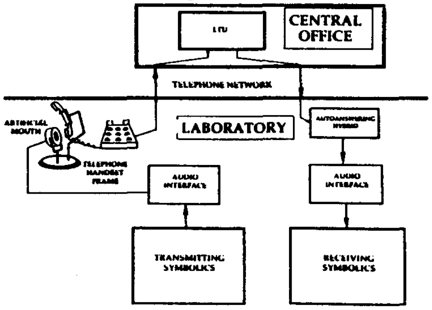

Figure 1 - Hardware Configuration for NTIMIT Transmission

For recording, an autoanswering "hybrid" device was used. This device automatically answers the telephone upon receiving a call, and electrically interfacs the telephone line with a standard audio connection.

To allow transmission of utterances to various locations, "loop-back" devices in remote central offices were used. These Line Test Units (LTU) have the capability to answer a telephone call over a normal telephone line, place a call to another specified telephone number, and link the incoming and outgoing audio paths. In this way, an utterance could be transmitted from the laboratory to a central office, then back from the central office to the laboratory. Fortunately, central offices with functioning LTUs existed throughout the NYNEX region, allowing utterance transmission to various locations. All LTUs in the NYNEX calling region were controlled by a central computer. This computer allowed choosing the LTU to be used for a particular audio path.

Two computers with two identical audio interfaces are used during the transmission procedure. The audio interface of the computer transmitting the TIMIT utterance is connected to the speaker on the artificial mouth, while the audio line on the hybrid device is connected to the audio interface of the recording computer.

The procedure for completing an audio path between the laboratory and a remote central office is as follows: First, the central computer is contacted from the laboratory and programmed to contact the desired LTU. The LTU is then called from the laboratory and the LTU "answers," opening half of the audio path. Through the central computer, the LTU is instructed to call another number in the laboratory. When the second number in the laboratory answers, the audio path is complete and transmission may begin.

4. Utterance Alignment Procedure

It is very desirable to have the phonetic and orthographic transcriptions in the TIMIT database reproduced in the NTIMIT database. A simple mapping of aligned transcriptions from TIMIT utterances to NTIMIT presented a problem, though, since the start times of TIMIT utterance playing and NTIMIT utterance recording could not be accurately controlled. Therefore, another method was used to align the NTIMIT utterances with the TIMIT utterances.

To facilitate alignment, sets of clicks were transmitted over the telephone network along with the TIMIT utterance. Two identical clicks were transmitted a known time period before the start of the TIMIT utterance, and two clicks were transmitted a known time after the end of the TIMIT utterance. The time between each of the two identical clicks is known.

After recording, the energy of the NTIMIT utterance is computed from 400 to 600 Hz every .5 milliseconds. Over the time region where it is suspected the clicks might be located, the following score is computed:

$$Score = \frac{E(n) + E(n+K)}{|E(n) - E(n+K)|} \qquad (1)$$

where E(n) is the energy value at sample n, and K is the known number of samples between the two clicks. The value of n that maximizes the score is considered the beginning sample of the first click. Knowledge of the time between the two clicks and the time between the clicks and the utterance allows calculation of the beginning time of the utterance. Using this information, the NTIMIT utterance can be edited automatically so that the beginning and end of the final NTIMIT utterance match the TIMIT utterance. In this way, simply copying the phonetic and orthographic transcriptions from the TIMIT utterance to the NTIMIT utterance will insure that the transcriptions are properly aligned.

This automatic procedure was successful approximately 90% of the time. Errors were identified by manually listening to the entire database. The usual result of an error was the presence of clicks in the final output utterance, indicating that clicks were not found in the time region specified by the automatic algorithm. The utterances with errors were retransmitted and realigned using the same automatic alignment procedure. Retransmitted utterances were sent to the same central office, but over a different telephone network path, since it was not possible to keep the same telephone line up for an extended period of time. Four passes of transmission were necessary to transmit the entire database.

5. Geographic Distribution of Utterances

The utterances in the NTIMIT database were transmitted over a variety of network conditions by varying the geographic location of the central office that the utterance was transmitted to. Given a rich set of LTUs and central offices which could be used for transmission, a procedure was developed to assign a central office to each TIMIT utterance. Not all available central offices were used, in order to speed the transmission process. The distribution process consisted of two stages: first each utterance was distributed to a broad geographic region, then the utterances are distributed to central offices within each region.

The broad geographic regions used correspond to Local Access and Transport Areas, or LATAs. LATAs are geographically contiguous areas corresponding to the subdivision of the telephone network. The NYNEX calling region of New York and most of New England consists of twelve LATAs, shown in figure [2]. Rhode Island, Vermont, New Hampshire, and Maine each are LATAs. Massachusetts and New York have multiple LATAs, roughly indicated by the dotted lines. A telephone call from a number in a given LATA to a number in the same LATA is considered "local"; a number call placed from one LATA to another is considered "long distance." NYNEX and its equivalent Regional Bell Operating Companies (RBOCs) across the United States handle local or intra-LATA telephone traffic, while long distance vendors such as AT&T, MCI, Sprint and others handle traffic between LATAs, or inter-LATA traffic. NYNEX and other RBOCs are also responsible for insuring that long distance vendors have adequate access to the local telephone network.

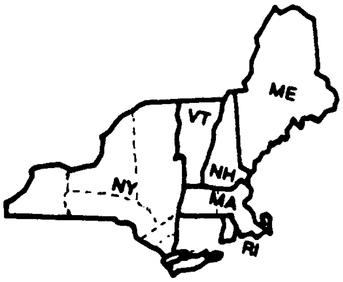

Figure 2 - LATAs in NYNEX calling area

Ten of the twelve NYNEX LATAs were used for TIMIT transmission. It was arbitrarily decided that half of the utterances would be transmitted "local," or within the New York Metropolitan LATA in which the NYNEX laboratory resides. The remaining 3150 utterances were distributed between the nine "long distance" LATAs. Table [1] shows how utterances were distributed to LATAs. In this table, each row corresponds to a speaker. The numbers in the table correspond to utterances for a given speaker. The full table would continue with this pattern for 630 speakers. This algorithm assured that each LATA would have approximately equal distributions of male and female speakers, and also various speaker dialects. These characteristics were verified after distributing the utterances to LATAs.

Once the utterances assigned to a LATA were determined, each utterance was distributed to a specific central office within the LATA. For a LATA with N central offices, the central office for utterance k within the LATA is simply k mod N.

After every utterance has been assigned to a central office, the transmission step begins. To streamline the process, transmission occurs one central office at a time. The audio path is opened using the procedure outlined in Section 3, then all of the utterances assigned to a central office are transmitted, along with the calibration tones described in the next section. The unaligned utterances are stored for later postprocessing, so that only transmission could occur during the day and alignment could proceed at night. Transmission required supervision to bring up the audio path and verify that the LTU in question was functioning properly.

The lack of correlation between LATA and line condition cannot be overstressed, since there was absolutely no way to control what kind of electrical path a given call would follow for any transmission. A call to a geographically distant LATA might have less distortion and noise than a local call, due to such effects as fiber-optic links along highly travelled long distance paths. Also, the rather complex system for routing calls most efficiently could easily route a rather short distance call through a distant city. This would degrade network conditions more than would be expected if one considered geographic distance alone. Since these same conditions would be expected under any telephone bandwidth application, this is an important factor to consider when differentiating between "local" and "long distance" network conditions, or attempting to associate geographic distance with line quality.

SPEAKER	LATA									
	1	2	3	4	5	6	7	8	9	10
1	01234	5	6	7	8	9				
2	90123		4	5	6	7	8			
3	89012			3	4	5	6	7		
4	78901				2	3	4	5	6	
5	67890					1	2	3	4	5
6	56789	4					0	1	2	3
7	45678	2	3					9	0	1
8	34567	0	1	2					8	9
9	23456	8	9	0	1					7
10	12345	6	7	8	9	0				
11	01234	5	6	7	8	9				
⋮					⋮					

Table 1 - Distribution of Utterances to LATAs

6. Calibration Levels, Signals and Tests

Several calibration measurements were made and test tones transmitted in order to accurately model the telephone transmission process. Also, tests were conducted to determine the effect of the acoustical modelling apparatus such as the artificial mouth and the telephone test frame.

Two levels required setting on the transmission hardware before transmission: the playback level of the transmitting audio interface and the record level of the receiving interface. The only consideration in setting the record level was to avoid peak clipping upon recording. Otherwise, this level was kept as high as possible to take advantage of the maximum possible dynamic range of the audio interface. The playback level on the output was slightly more involved. It was important to have the sound level at the output of the artificial mouth at "typical" levels for an average telephone conversation. An abnormally high sound level could overdrive the telephone microphone, introducing extraneous harmonic distortion. Higher than average sound levels would also create a higher than normal signal to noise ratio on the telephone line. Both of these effects greatly impair the accuracy of the artificial mouth/test frame simulation.

To set this level, we used a standard set by operator headset designers; the sound level at the microphone of a headset (or handset in this case) should be 88 dB SPL. Informal listening tests with "average" TIMIT utterances verified that this seemed a reasonable standard for setting the playback level, so this standard was adhered to throughout the transmission sessions.

For each central office used during transmission, two calibration tones were sent along with the TIMIT utterances. These tones were a 1 KHz, 5 second tone and a 0-4 KHz, 2 second linear frequency sweep tone. Using these tones, network characteristics such as attenuation, frequency response, and harmonic distortion can be readily determined for a given transmission path. Figure [3] shows a Fourier Transform of a 1 KHz tone from a "local" central office. Figure [4] shows a frequency response derived from the time dependent total energy of the sweep tone from the same central office. This shows the frequency response of that particular line, while the 1 KHz tone shows harmonic distortion at that frequency. The jaggedness in the frequency response is due to the fact that the total energy picks up the energy in the the harmonics as well as the fundamental frequency, slightly skewing the results. Future studies will compare the calibration tones from various central offices to determine how they effect the characteristics of the telephone line.

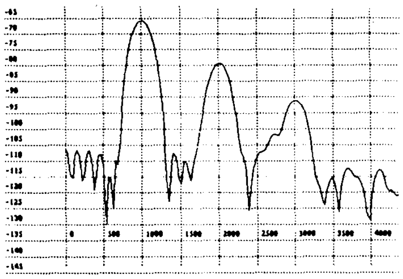

Figure 3 - 1 KHz Tone From "Local" Central Office

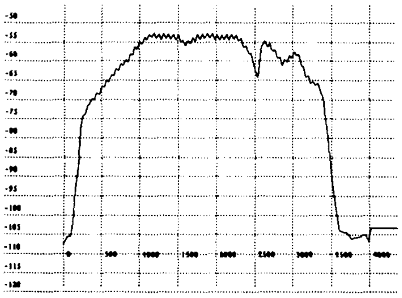

Figure 4 - Frequency Response From Total Energy

Finally, an experiment was conducted to determine the accuracy of the "simulation." The ideal NTIMIT database would be collected by rerecording all of the TIMIT sentences using a normal telephone. Since this is not possible, the artificial mouth and telephone test frame were used to approximate the acoustic properties of a mouth-telephone coupling. An important test for the simulation is the difference between the ideal system and the approximation used in the NTIMIT transfer process.

This test was conducted as follows: one male speaker recorded a set of randomly chosen TIMIT sentences through both the NTIMIT telephone handset and the same model of microphone headset used in recording the TIMIT database. This results in one wideband and one narrowband version of the sentences. Keeping the telephone line up, the wideband version is then transmitted over the telephone network using the NTIMIT hardware setup. In this way, there are now two narrowband versions of the sentences; one transmitted directly over the telephone, and the second indirectly transmitted through the artificial mouth and handset test frame. The direct version is the ideal NTIMIT recording scheme; the indirect version is created using the same method used for collecting NTIMIT. Relevant differences between these two signals reflect distortions caused by the artificial mouth and telephone test frame setup.

Both narrowband version were manually aligned, and segmented into speech segments, eliminating large portions of silence. A 256 point wideband short-time Fourier Transform was computed for both of the sets of speech segments, and the average difference between the sets computed. A graph of this difference appears in Figure 5. In the range of interest for telephone bandwidth, i.e. 0 Hz - 3500 Hz, we see a broad drop of approximately 3 dB spanning from 2000 Hz to 3500 Hz. It is not currently known what the source of this difference is. No efforts have been taken yet to correct for this difference in the NTIMIT data, although corrections are under consideration.

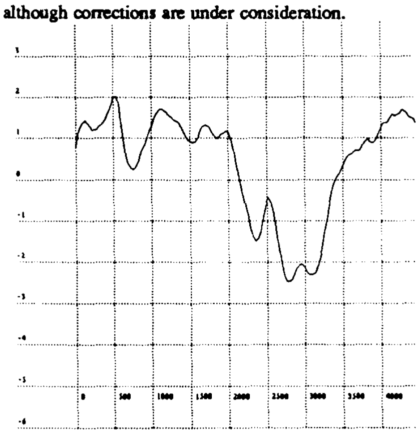

Figure 5 - Frequency Response of Simulation Error

7. Conclusion

The collection of a phonetically balanced, multi-speaker, continuous speech, telephone bandwidth speech database has been described. This database has many useful characteristics of general speech databases, plus the properties introduced by the telephone network. This is useful for telephone bandwidth speech recognition algorithms, or analysis of telephone bandwidth speech.

REFERENCES

[1] William M. Fisher, George R. Doddington, and Kathleen M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proc. DARPA Workshop on Speech Recognition*, pp. 93-99, February 1986.

[2] Stephanie Seneff and Victor W. Zue, "Transcription and Alignment of the TIMIT Database," 1988.