Analysis of Two Algorithms for Telephone Speech Recognition
Benjamin Chigier, Erik Urdang, Judith Spitz
NYNEX Science and Technology

ABSTRACT

The telephone network presents speech recognition devices with a band-limited, noisy, and, in some cases, distorted speech signal. A series of experiments were performed to quantify the effects of these transformations on two current recognition algorithms: a) an acoustic segmentation algorithm and b) an acoustic classification algorithm. The data used in these experiments are a subset of the TIMIT speech database and a telephone network version of the identical TIMIT utterances (N-TIMIT). In this paper, we present insertion and deletion results for the segmenter (for both conditions, compared to hand transcriptions) as well as patterns observed in segmentation errors as a function of data set. Also presented will be the results of the classification algorithm for both databases.

Introduction

The telephone network has many effects on the speech signal it carries. Some of the more obvious changes are: band limitation (below approximately 300 Hz and above approximately 3400 Hz), addition of noise (which may vary widely over different connections and switches, even within a given call), echo, ringing, and crosstalk. The following figure shows some of the effects of the network.

Although these effects have been well quantified, there has not been systematic investigation of the effects these changes have on the performance of speech recognition algorithms. The goal of this pilot study was to examine the effect of the telephone network transmission channel on the performance of two recognition algorithms:

- An acoustic segmenter, and
- An acoustic labeller.

It should be noted that the goal was not to attribute particular errors to specific properties of the telephone network, but rather to identify and characterize the differences between the algorithms' performance on speech given the two conditions (network vs. broadband).

Experimental Design

Stimuli

The speech used in this experiment consists of sentences excerpted from the TIMIT database [1, 4]. The TIMIT database consists of 10 sentences spoken by each of 630 talkers. The 10 sentences may be broken down as follows: 2 calibration sentences (the same 2 sentences were used for all talkers) for the purpose of evaluating regional dialects, 3 sentences selected at random from the Brown Corpus [3], and 5 sentences designed at MIT to provide as wide as possible a coverage of phoneme pairs.

For the purpose of this study, 355 utterances were selected from the database. More specifically, 71 speakers were randomly selected and their 5 MIT sentences were collected. A telephone network version of each of the 355 utterances was created as described below.

N-TIMIT

A telephone version of these utterances was created to form a "network" database (hereafter referred to as N-TIMIT) [5]. In order to accomplish this, the following equipment was used:

- 2 Symbolics 3650 Lisp Machines
- 2 Digital Sound Corporation DA/AD converters (model #200)
- A Briel & Kjaer Mouth Simulator Type 4227
- An IAC sound isolation booth
- A standard telephone
- A variable telephone network connection and interface.

The two Symbolics' were employed simultaneously, one to transmit the utterances (Symbolics A) and the other to receive them (Symbolics B). Each of these was connected to a DSC 200 (see Figure 1). The Mouth Simulator and telephone were placed together in the sound isolation booth. Given this configuration, the transmission path for a given digitized utterance was as follows: Symbolics A sent the utterance through the DSC which converted it to an analog waveform and sent the signal to the Mouth Simulator in the sound isolation booth. The utterance was then "spoken" by the Mouth Simulator and picked up by the telephone microphone (which was held at a fixed distance and angle from the Mouth Simulator) and was then transmitted through the telephone network (described below), digitized by the second DSC 200 and stored by Symbolics B. Once the utterance was fully transmitted, Symbolics A then notified Symbolics B that it should stop recording, and get ready to accept the next transmission.
Results and Analysis

Segmentation Inferences

A Log Linear Analysis was performed on the number of insertions made by the segmenter as a function of database (TIMIT vs. N-TIMIT) and the context in which each insertion could have occurred. This analysis revealed a significant effect of database on insertions (p < .001). More specifically, a greater number of insertions were found in the N-TIMIT condition than in the TIMIT condition (788 vs. 578, respectively). In addition, for each database, the analysis revealed a significant effect of context on the number of insertions made (p < .001) and a significant interaction between context and database (p < .001).
Due to the variance in the number of instances in each context, it is difficult to state precisely how each context contributed to this overall result, but the data strongly suggest that the N-TIMIT database was more likely than the TIMIT database to have insertions in the contexts Vowel_Fricative and Fricative_Vowel.

**Segmentation Deletions**

A Log Linear Analysis was performed on the number of deletions made as a function of database, context and target. The results again revealed a significant effect of database on the number of deletions: 487 deleted in N-TIMIT, 405 deleted in TIMIT (p < .001). There was found to be no significant effect of context on database. On the other hand, a significant interaction was found between target and database (p < .001).

As with insertions, due to the variance in the numbers of instances across target categories, it is difficult to say with certainty which targets were strictly responsible for the interaction, but a large number of the deletions seemed to be caused by alveolar stops (both closures and releases) which may be explained by the fact that much of the information that distinguishes the closure from the burst in an alveolar stop lies above 3400 Hz.

**Labelling**

In order to determine whether there was a significant effect of database on labelling accuracy, a McNemar Test of Symmetry was conducted for each of the seven labels, as well as one for the set of all labels. The results of this analysis revealed a significant effect of database (TIMIT vs. N-TIMIT) for every label except /a/ and /l/. That is, labelling accuracy was superior for the TIMIT utterances than for the N-TIMIT utterances. Table 2 shows these results (see Note [1]):

<table>
<thead>
<tr>
<th>Label</th>
<th>Probability</th>
<th>Number Tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>/a/</td>
<td>0.0833</td>
<td>57</td>
</tr>
<tr>
<td>/i/</td>
<td>0.0831</td>
<td>74</td>
</tr>
<tr>
<td>/u/</td>
<td>&lt; .001</td>
<td>89</td>
</tr>
<tr>
<td>/o/</td>
<td>&lt; .001</td>
<td>87</td>
</tr>
<tr>
<td>/l/</td>
<td>&lt; .001</td>
<td>131</td>
</tr>
<tr>
<td>/l/</td>
<td>&lt; .001</td>
<td>125</td>
</tr>
<tr>
<td>/l/</td>
<td>N/A</td>
<td>33</td>
</tr>
<tr>
<td>All</td>
<td>0.0000</td>
<td>596</td>
</tr>
</tbody>
</table>

**Conclusions**

This pilot research has shown that the telephone network has a significant effect on the performance of recognition algorithms. In addition, it has provided a methodology for associating particular effects with specific phonetic categories. In order to build a successful telephone channel speech recognizer, it is crucial to understand these effects as a preliminary to developing "network tolerant" algorithms.

**Future Work**

It has become apparent that, in order to further analyze these results, it will be necessary to gather data on considerably more utterances (thousands). By doing this, we hope to be able to report, in much greater detail, on the specific classes of phonemes likely to be more difficult for a telephone network recognizer than for a broadband recognizer. In addition to this we plan to use the techniques described in this paper to evaluate the efficacy of various signal representations with respect to communication channel. In essence, these techniques will enable us to evaluate the effects of various front ends on classification and segmentation in order to pick representations that optimize particular aspects of recognizer performance.

**Notes**

[1] There is no entry shown for label /2/: this is because there were no instances where the labeller correctly identified a token in this category in the N-TIMIT condition.