DISCRIMINATIVELY DERIVED HMM-BASED ANNOUNCEMENT MODELING APPROACH FOR NOISE CONTROL AVOIDING THE PROBLEM OF FALSE ALARMS.

R. Chenganvarayan and D.L. Thomson
Speech Processing Group, Lucent Speech Solutions Department
Lucent Technologies, Naperville, IL 60566, USA
Email: ruth@lucent.com, davidt@lucent.com

ABSTRACT

Earlier we proposed modeling echo residuals by using multiple echo models built from a set of specific announcement. Experienced callers may interrupt the prompt by speaking the keywords over the prompt. This leads to incomplete prompt echoes that were not modeled by multiple echo models. In this study, we investigate further improvements by building an echo model of each word in the entire announcement, then linking each model in sequence to track the exact echo that precedes valid speech (movie title). The experimental results show that by modeling exactly, one can get better recognition accuracy and less false triggering, with a possible increase in computational complexity.

1. INTRODUCTION

As the demand for automatic speech recognition (ASR) technologies keeps on the rise, the need for the development of systems that are robust to speaking style, accents, environmental mismatch etc. is becoming increasingly important [8, 9]. In these situations, integrated echo and filler models play an important role in maintaining an acceptable error rate in providing a desirable trade-off between false alarm and false rejection rate [1, 5, 12, 11]. In mobile telephony, speech signals are often degraded by the presence of acoustic background noise as well as by system introduced interference [6]. This has an adverse effect on the perceived quality and intelligibility of speech as well as on the performance of speech processing applications in the network [4].

When the interference is an echo of the system announcement, then echo cancellers are able to reduce the echo by up to about 25 dB and they generate very few artifacts [6]. However, if the echo is loud and the incoming speech is quiet, the residual echo energy following cancellation still begins to approach that of the incoming speech. Even when these echoes do not trigger ASR, they interfere with recognition of valid input. Recently, we proposed building a model for the echo residual so that the recognizer is able to identify that the incoming signal is an echo, not speech from a caller [4]. The narrow training strategy based on single recorded prompt made the echo model a better discriminator, though it has the possible drawback of requiring retraining if the prompt changes. To overcome this intrinsic problem, we described building a more general echo model that is trained over many sentences from varied voices [1]. This ability makes the recognizer more resistant to interfering residual echoes.

New users typically listen to the entire system prompt prior to responding, whereas experienced callers may interrupt the announcement by speaking the keywords over the prompt. This leads to incomplete prompt echoes that were not properly modeled by a single or multiple prompt-based echo models. In this paper, we explore further improvements by building an echo model of each word in the entire announcement, then linking each model in sequence to track the exact echo that precedes valid keyword speech (movie title). We would allow arcs from the end of each

Figure 1. Energy measurement contours and input signal for the movie title "Shakespeare in Love" spoken by a male speaker. The data was collected over a T1 (digital line) and contains no echoes.

2. NETWORK ECHO CANCELLATION

In a pure digital network there is little or no echo as illustrated in Figure 1. Often, however, the analog portion of the voice path creates echoes due to impedance mismatches at 2-wire to 4-wire hybrids. Echoes can also be caused by acoustic coupling...
between the microphones and loudspeakers as in speakerphones [6]. The system prompt as shown in Figure 2 is fed to the end- user and he or she, in turn, speaks back to the ASR system. The user input is corrupted with an additive echo that results from the reflection of the system prompt due to the presence of one or more hybrids as shown in Figure 3. This resulting echo needs to be cancelled prior to performing ASR. Not doing echo cancellation prior to ASR would result in the system prompt falsely triggering the recognition system. The echo canceller takes the system prompt and the echo corrupted user speech as input and uses an adaptive finite impulse response (FIR) filter to remove as much as the echo as possible [6]. The resultant echo cancelled speech output from the echo canceller is shown in Figure 4. This is then fed as input to the ASR system for further processing as depicted in Figure 5.

3. ANNOUNCEMENT MODELING

If the echo is loud and the incoming speech is quiet, then the residual echo following even very good echo cancellation begins to approach the magnitude of the incoming speech. These echo residuals can interfere with proper recognition. In this paper, we propose to model the echoes using hidden Markov model (HMM) based maximum likelihood (MLE) and minimum classification error (MCE) training methods. By building an echo model of each word in the entire announcement, then linking each model in sequence to track the exact echo that precedes valid speech the recognizer is able to identify that the incoming signal is an echo, not speech from a caller. This ability makes the recognizer more resistant to any kind of interfering residual echoes.

We illustrate this technique using a movie title recognition task. We based the echo model on recordings of users calling a system that provides movie information. Of these recordings, 1313 speech files were manually selected that contained only residual echoes (no speech from callers) of the prompt. These files had been processed by the echo canceller and represented a variety of acoustic environments with varied signal strengths and durations. The specific prompt was “Please say the name of the movie you would like to see,” recorded by a male speaker. This prompt is modeled by six concatenated word models as shown in Figure 7. The six word strings are please, say the, name of, the...
movie, you would like to see. We built six different word models, each with a 10-state, 4-mixture continuous density HMM using one iteration of MLE and five iterations of MCE training using all the available echo files [2]. Only the silence and the echo word model parameters were updated during MCE training and all the other speech models were untouched.

The grammar main network and the corresponding sub-networks for the recognition task are shown in Figures 6 and 7. This framework models the prompt echoes via the echo model positioned in the front of movie title. The incoming speech can exit in either the silence/filler module or the echo model and may loop back and forth between these two models until a valid movie title segment is found. Make a note that once the decoder leaves a particular echo word then it cannot come back to the same word model again for that particular utterance. The decoder has to follow either the next word or it can stay in silence/filler module or it can go directly to the keyword recognition module. This constraint allows the recognizer to follow a left-to-right pattern since the words in the prompt are spoken in sequential manner. For example, the given speech utterance may contain few echo words and any one of the keywords. In this case, the decoder stays in those few echo words and then it may stay in silence/filler module for a while until the valid keyword segment is reached. Then backward phone models with grammar constraints are used to decode the spoken keyword (movie title). The advantage of the approach in Figure 6 is that any complete or partially prompted announcement echo residual can be absorbed by the concatenated word-based echo models.

The residual output from an echo canceller has a unique signature that is characterized by the echo model. For example, it tends to be uniformly whiter than normal speech. This distinction allows the echo model to effectively help the recognizer distinguish echoes from caller input. Although one can consider building a general echo model, trained over many sentences from varied prompts, the experiment here used an echo model based on single recorded prompt. This narrow training strategy makes the echo model a better discriminator, though it has the possible drawback of requiring retraining if the prompt changes. This problem can be solved by building a more general echo model that is trained over many different prompts from varied voices as described in our recent study [1]. We have not explored this alternative in this work.
4. EXPERIMENTAL RESULTS

The task was a trial of an ASR based service which accepted spoken queries from customers over the telephone concerning movies playing at local theaters. This task, referred to as the movie list, was interesting for evaluating noise control techniques because it contained a large number of ill-formed utterances, speakers with a variety of speaking styles and energy levels, and relatively strong announcement echoes. The movie list testing set consisted of 817 speech files containing 66 current movie titles (Broken vessels, Wild wild west, Austin powers etc., featured during the summer of 1999) spoken by about 12 native speakers. Due to the nature of the analog portion of the telephone connection in the trial, echoes were significantly stronger than those encountered in a typical speech recognition system, thus presenting the recognizer and echo canceller with an unusually difficult task. The training sets for building the context-dependent subword models consisted of 9685 generic English phrases recorded over the telephone network in a U.S. wide data collection covering each dialect region.

Input speech is sampled at 8kHz and segmented into overlapping frames 30 msec long with centers 10 msec apart. Each frame is windowed with a Hamming window and then processed using a 10th-order LPC Analyzer. The recognizer feature set consists of 39 features that include the 12 LPC-Gerived lifter cepstral coefficients, log-energy and their first and second order derivatives [2]. The energy feature is batch normalized during training and testing [3]. Since the signal was recorded under various telephone conditions and with different transducer equipment, each cepstral vector was further processed using batch-mode cepstral mean subtraction to reduce the effect of channel distortion [3].

The subword model set used in recognition consists of 41 context independent units [2]. Each subword is modeled by a three-state left-to-right continuous density HMM with only self and forward transitions. A mixture of Gaussian functions with diagonal covariances is employed to estimate the density function for each state. A maximum of 16 mixtures per state is allowed. The silence/background is modeled with a single state, 32 Gaussian mixture HMM [2]. The filler or garbage is modeled with a single state, 64 Gaussian mixture HMM [1]. Lexical representations of movie titles are obtained by preprocessing the orthographic transcriptions through a pronunciation software module designed for use in text-to-speech. The initial model set of 41 subword units is trained using a conventional maximum likelihood training procedure [3]. We then apply five iterations of MCE training to the initial boot model with null grammar. The number of competing string models is set to four [7].

The movie list speech recognition results showing comparative performance of the baseline and baseline with echo model are summarized in Table 1. The results showed that the multiple echo models built for a set of specific announcement provide good accuracy and much echo-resistance. We observed a 32% string error rate reduction by using the multiple echo residual models compared to using the echo canceller alone. Most of the false triggers due to strong echoes are removed when using the proposed echo modeling approach. To evaluate out-of-vocabulary performance, a database of 6000 utterances consisting of short non-digit phrases spoken by a variety of speakers was selected and an unknown length grammar was used. When the recognizer was presented with short non-digit (out-of-vocabulary) phrases, it did not report recognition of keywords on roughly 85% of the sentences and wrongly triggered on the remaining 15%.

This is still substantially better than a baseline system which would have wrongly triggered on 100% of the non-digit phrases. By augmenting the multiple echo models in parallel with filler models further lowered the rate of incorrect recognition (insertion error) when out-of-vocabulary speech is encountered.

5. CONCLUSION

In this work, we proposed modeling echo residuals using HMM-based maximum-likelihood and minimum classification error training methods. This approach proved to be effective in eliminating false triggering and improving recognition accuracy. Test results showed that multiple word-based echo models built for a set of specific announcement provided good accuracy. Most false triggers due to strong echoes were eliminated. The experimental results further demonstrated that the multiple word-based echo models in conjunction with suitable filler model to represent the extraneous speech not only provide good recognition accuracy but also yield better out-of-vocabulary rejection.

REFERENCES


