Abstract: This paper describes the LDC forced aligner which is designed to align audio and transcripts. Unlike existing forced aligners, LDC forced aligner can align partially transcribed audio files, and also audio files with large chunks of non-speech segments, such as noise, music, silence etc., by inserting optional wildcard phoneme sequences between sentence or paragraph boundaries. Based on the HTK tool kit, LDC forced aligner can align audio and transcript on sentence or word level. This paper also reports its usage on English and Mandarin Chinese data.

Introduction

Purpose: align audio to incomplete transcript

Existing forced aligners require
• audio files have been completely transcribed
• there are no long periods of non-speech regions, such as silence, music, background noise

LDC forced aligner
• can align partially transcribed audio files
• uses wildcard phoneme sequences to match audio segments without transcripts
• is based on HTK Tool Kit

Alignment

LDC’s FA doesn’t use HTK’s forced alignment feature
Alignment comes as a byproduct of recognition
Insert wildcard phoneme sequence in the grammar:

sentence1 {WILDCARD1 | WILDCARD2 |…|WILDCARD25}
sentence2 {WILDCARD1 | WILDCARD2 |…|WILDCARD25}
sentence3 {WILDCARD1 | WILDCARD2 |…|WILDCARD25}

Creation of Wildcard Sequences

24 wildcard sequences were created as follows
① recognition system on broadcast news data
• speech with or without non-speech background
• Non-speech segments, such as coughing, laughing, music and other noises, were excluded from the training data;
② Divide development data into four categories: speech, music, noise, and speech with non-speech background;
③ run the speech recognizer obtained in 1) on each category of 2);
④ select the most frequent six triphones from each of the four categories;
Silence as the 25th wildcard

Implementation

Two systems are implemented:
• English
• Mandarin Chinese

English system
• Trained on 67 hours of broadcast news audio
• Used CMU pronunciation dictionary
• 125,000 words
• 36 phonemes
• OOV words
• Transducer trained on CMUdict to create pronunciation
• Numbers, symbols, acronyms
• rule-based system to spell out numbers and symbols
• all possible readings provided

Chinese system
• Trained on 63 hours of broadcast news audio
• Used a Pinyin lookup table as the pronunciation dictionary
• 7,333 Chinese characters
• 38 phonemes
• No OOV words

Experiments

Test data
• 30 hours of broadcast news per language
• Transcript paragraphs randomly removed
• Audio file length ranges between 10 to 60 minutes
• 10.5 to 35.8 percent of audio don’t have transcripts
• Forced alignment on full transcripts used as gold standard

Results

English
• average word level distance from gold standard: 10ms
• average sentence level distance from gold standard: 15ms

Chinese
• average word level distance from gold standard: 12ms
• average sentence level distance from gold standard: 18ms