Automatic Transcription of Continuous Speech using Unsupervised and Incremental Training

G.L. Sarada, N. Hemalatha, T. Nagarajan,
Hema A. Murthy

Department of Computer Science and Engg.,
IIT, Madras.

INTERSPEECH-2004
POSTER PRESENTATION

Abstract

- Conventional way of speech transcription
  - Bootstrapping
    - Existing speech recognizer used to transcribe the new data
    - Needs manually transcribed data
- Challenges
  - In a country like India
    - 22 official and nearly 5000 unofficial languages
    - Need for large amounts of transcribed data for speech recognizers

Need for Automatic Transcription system with minimum amount of manual work

- Automatic Transcription Using Unsupervised and Incremental Clustering Technique

  Involves

  - Automatic Segmentation
  - Unsupervised and Incremental Training Technique[†]
  - Labeling

- Addressed the issues in the baseline system and made several refinements to it

- Obtained performance improvement of 8% over the baseline system

This is compared to the baseline system

Introduction

- Today's state-of-the-art SR systems are able to transcribe unrestricted broadcast news with good accuracy

  Issues:

  - Relies on the large amounts of manually transcribed training data
  - Obtaining such data is time consuming and expensive
  - Requires trained human annotators and substantial amounts of super vision

- To overcome above problems, most commonly used methods are

  1. Bootstrapping
     - Recognizer trained with 1hr of manually transcribed speech used for transcribing the rest of the data
     - Again used to train the recognizer
  2. Automatic segmentation and labeling when it's orthographic projection is given

Issues:

- Poor performance due to mismatch of environment or language
- Slow convergence during refinement of models

- Novel approach for automatic segmentation and transcription of speech data without using manually annotated speech corpora
  - Speech segmented into syllable-like units
  - Incremental Training

  Issues:

  1. Poor clustering due to syllable segments/merged syllables
     eg1 : /**vana**/ having two vowels and consonants cluster with other syllables having similar **V**/**C** part
     eg2 : /**k**/ having short duration segment
  2. Clustering is poor because of syllable segments having the silence at the boundaries

- Above mentioned approach is used as the baseline system
  - Refinements are made to overcome the problems in incremental training

Syllable-like Segmentation

- As Indian languages are syllable-timed, speech data is segmented into syllable-like units
  - Group delay based automatic segmentation into syllable-like units
  - Processing of Short term energy of the speech signal
  - The group delay spectrum is obtained from the inverted short time energy
    - Peaks are extracted
    - Location of peaks corresponds to the syllable boundaries
  - Prepend and append small duration silence to the syllables

Initial Cluster Selection

- Incremental training leads to fast convergence if similar syllables in each cluster
  - Take All $\mathcal{N}$ syllable segments for initialization of models.
  - Extraction of features (13 MFCC + 13 delta + 13 acceleration) with multiple resolutions
    - Ensures a reasonable variance for each Gaussian mixture in the models.
  - Initialization of $\mathcal{N}$ Hidden Markov Models
  - $\mathcal{N}$ syllable segments are decoded using 2-best criteria.
    - Results in $\mathcal{N}$ pairs of syllable segments
  - Pruning # of models based on the repetition of the syllable segments.
  - Create new models with reduced # of pairs.
  - Repeat above steps for **m** times
    Leads to $\mathcal{N}1$ clusters where $\mathcal{N}1<\mathcal{N}$which have similar syllable segments.

Incremental Training

- Steps followed in incremental training
  1. Re-estimation of model parameters using Baum-Welch re-estimation
     - Each model is a 7 state 1 Gaussian mixture HMMs.
  2. New models are used to decode all the syllable segments using Viterbi decoding.
  3. Clustering based on the decoded sequence.
  4. Reduction in # of clusters based on # of syllable segments in them
  5. Repeat steps 1-3 until convergence is met

Flow chart: Unsupervised and Incremental Training

Convergence Criteria

- Re-estimation of model parameters and re-clustering of syllable segments
- Reduction in # of syllable migrations from one cluster to another
- Convergence is met when # of migrations becomes zero
- Terminate incremental training procedure.
- Produces $\mathcal{N}2$ ($\mathcal{N}2 < \mathcal{N}1$) syllable clusters

Identical/similar syllable segments in each cluster, with a few exceptions.

Labeling Clusters and Transcription

- Required to assign a label for each of the clusters for transcription/recognition tasks.
- Manual labeling
- Use models with labels for transcription/recognition of speech data.

Performance analysis - (a) an example of speech signal. (b) Group delay spectrum of the speech signal. A.Trans - Automatic transcription. M.Trans - Manual Transcription.

Performance Analysis

- Four female speakers data each of 15min duration for training the system
- During testing, two kinds of data:
  - Untranscribed data corresponding to speaker used in training.
  - Untranscribed data corresponding to speaker not used in training.
- Prepend and append short duration silence of $\approx 20ms$ to the syllable segments
- Obtained performance improvement of 15% for I and 8% for II as a syllable recognizer
- Obtained performance improvement of 22% and 12% as a $\mathbf{CV}/\mathbf{VC}$ unit recognizer
- Considerable reduction in the performance for False case

Table 1: *Performance (in %) analysis of baseline system before refinement and after refinement*

Sound units	Before refinement		After refinement	
	I	II	I	II
Syllables	41.98	34.98	56.2	42.6
CV+VC	18.52	16.7	25.6	20.8
Vowel only	27.30	31.0	13	27.2
Cons. only	3.25	4.285	2.4	3
False	8.95	13.03	2.8	6.4

Conclusions

- Refined the base-line system to improve the performance of the transcription system which segments and transcribes the continuous speech signal without the benefit of manually annotated speech corpus.

- Obtained performance of 56% and 42% for known and unknown speaker data respectively.

References

(1) Nagarajan., T. and Murthy., H. A., "An approach to segmentation and Labeling of continuous speech without bootstrapping", NCC-2004, pp.508-512, Jan 2004.

(2) Frank Wessel and Herman Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition", IEEE workshop on ASRU, pp.307-310, Dec 2001.

(3) Ljolje., A. and Riley., M. D., "Automatic segmentation and labeling of speech", ICASSP-1991,Vol.1, pp.473-476, April 1991.

(4) Lori Lamel, Jean-Luc Gauvain and Gilles Adda, "Unsupervised acoustic model training", ICASSP-2002, Vol.1, pp.877-880. May 2002.

(5) Shuangyu Chang, Lokendra Shastri and Steven Greenberg, "Automatic phonetic transcription of spontaneous speech, (American English)", ICSLP - 2000, Vol.4, pp.330-333.

(6) Database for Indian Languages, India, speech and vision laboratory, IIT Madras, Chennai-2001.

(7) Prasad., K. V., Nagarajan., T. and Murthy., H. A., "Automatic segmentation of continuous speech using minimum phase group delay functions, Speech Communications, Vol.42, pp.429-446, April 2004.