Abstract

Content indexing has become necessary, not just optional, in the era where broadcast, cable and Internet produce huge amounts of media daily. Text information from spoken audio is still a key feature to understand content along with other meta-data and video features. In this paper, a new method is introduced to improve transcription quality, which allows more accurate content indexing. Our method finds phonetic similarities between two imperfect sources, closed captions and ASR outputs, and aligns them together to make quality transcriptions. In the process, even out-of-vocabulary words could be learned automatically. Given broadcast news audio and closed captions, our experimental results show that the proposed method, on average, improves word correct rates 11% from the ASR output using the baseline language model and 6% from the one using the adapted language model.

Index Terms: Content indexing, Closed captions, Phonetic alignment

1. Introduction

Content creators face a number of challenges in delivering their product to the largest possible audience. Distributions systems may include broadcast, cable, or IPTV television services, and there is a growing number of Internet and mobile devices that may require customized video encodings to create the best user experience. Services like AT&T’s Internet Video Management Service (iVMS) offer end-to-end solutions with advanced content distribution networks, reporting, and optimized media transcoding [1].

Next generation services include content indexing capabilities such as story segmentation with thumbnail generation and automatic speech recognition (ASR) to help content creators manage large collections and allow users to locate desired content. By creating content indices with temporal information, these services can identify points of interest in long form content in response to user queries and stream the media accordingly.

Transcription of spoken audio is an important feature for content indexing since it carries most of the significant information among other features produced from content streams. Even though video processing techniques have evolved greatly, spoken text from the audio stream still carries more semantic information to understand content easily. Typically, there are two kinds of transcriptions available, 1) closed caption and 2) ASR output, derived from the broadcast content.

Many television programs include closed captions appearing at the bottom of the screen to assist the hearing impaired. Closed captions produced by human transcribers are very accurate and also reliable to catch names and new terms. However, they suffer lag problems constraining its application in video indexing and drop words especially during periods of rapid dialog. On the other hand, ASR output synchronizes in time with the actual utterance, but even the state-of-the-art ASR systems cannot recognize out-of-vocabulary words, especially foreign names or new terms, which would be critical for content indexing.

There have been many efforts to improve the performance of broadcast news transcription by ASR acoustic model training [2] [3] [4], language model adaptation [5] [6] [7], hypothesis re-ranking [8], and text-based closed caption alignment [9]. However, previous works which try to align captions and ASR outputs at the word-level suffer from ASR errors, because ASR systems tend to produce erroneous words when they face out-of-vocabulary words or background noise.

In this paper, a new method is introduced to phonetically align two imperfect sources of transcription, closed caption and ASR output and improve the quality of ASR transcription. When words in captions and ones in ASR outputs don’t match at the orthography level, the new method converts the words into a sequence of phoneme strings and finds similarities on both sides. It is observed that erroneous words generated by the ASR system don’t make sense by themselves, but their pronunciations were similar to the words spoken originally.

2. System Overview

The AT&T MIRACLE (Multimedia Information Retrieval by Content) system has been archiving and indexing multiple channels of broadcast news and programs daily over the years [10]. This system as shown in Fig 1 can process media from a wide range of digital video sources, but for this work, broadcast news and other television programs were acquired from a consumer-grade Digital Broadcast Satellite receiver, converted into analog and encoded into MPEG-2 format using a low-cost PC card.
2.1. Corpus

The corpus used in this work consists of 21 sessions which are approximately 30-minute ABC and NBC evening news programs in a two week period in February 2007. The audio is extracted from 48kHz Stereo MPEG audio encoded with MPEG 1 Audio Layer III at 256Kb/s, transcoded into 16 bit linear PCM and down-sampled to 16kHz for further processing including ASR. The authors asked the AT&T Labs annotation group to transcribe the collected audio in the same format as that of the 1996 CSR Hub-4 evaluation corpus.

2.2. Closed Caption

To comply with FCC accessibility mandates, most television programs include closed captions for the hearing impaired. Closed captions, which were manually transcribed by human experts, are clearly one of the most important sources for content indexing services.

Two methods used to generate captions are off-line and real-time. In this work, real-time captions are processed for prompt content indexing since they are aired with live broadcast news. Real-time captioning, however, introduces a several second or longer delay because even the fastest human transcribers need to first hear the speech before typing it out. This delay would cause problems for content search since the index would not be accurate.

![Figure 2: Distribution of closed caption delay](image)

Fig 2 shows the lagging-time distribution of closed captions in the collected live broadcast news data set. The delay of the caption varies from 2 to 8 seconds.

Another problem in real-time closed captions is the lack of accuracy. Even if closed captions are transcribed by a human transcriber, they have lots of missed words or may contain typos because the human transcriber can’t keep up with the spoken audio.

Table 1: Closed caption accuracy in live broadcast news (WNT: World News Tonight, NN: Nightly News)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Insertion</th>
<th>Deletion</th>
<th>Substitution</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNT</td>
<td>1.0%</td>
<td>20.1%</td>
<td>4.0%</td>
<td>75.9%</td>
</tr>
<tr>
<td>NN</td>
<td>1.0%</td>
<td>16.8%</td>
<td>5.9%</td>
<td>77.4%</td>
</tr>
</tbody>
</table>

The closed captions in our broadcast news corpus contains around 18 to 21% word-error rate (WER) including mainly deletions and some substitutions as shown in Table 1. Some of the substitution errors contain words which have literally equivalent meanings but which are spelled differently, such as did agree → agreed, no one → nobody, and multi millionaire → rich, etc.

2.3. ASR System Outputs

Another source of extracting transcription of spoken audio streams is to utilize automatic speech recognition (ASR) technology. Thanks to the evolution of acoustic and language model training, ASR systems provide nearly instantaneous transcription but ASR system outputs become less accurate when audio is degraded due to the presence of background noise, or when a person in the content says less frequently occurring words such as foreign names.

To quantify our ASR performance degradation in different acoustic conditions, we partitioned the broadcast news audio with the following conditions which were used at 1996 CSR Hub-4 evaluation.

- **F0** Baseline Broadcast Speech
- **F1** Spontaneous Broadcast Speech
- **F2** Speech Over Telephone Channels
- **F3** Speech in the Presence of Background Music
- **F4** Speech Under Degraded Acoustic Conditions
- **F5** Speech from Non-Native Speakers
- **Fx** All Other Combinations

As shown Table 2, accuracies of closed captions are fairly consistent in different acoustic conditions while ones from ASR are significantly lower in the acoustically-degraded conditions. Our ASR system outperforms a human captioner when it processes speech recorded in the studio with the planned script (F0). In all other conditions, however, closed captions by a human annotator were more reliable than ASR outputs, except speech in background music (F3) where anchor persons’ names are skipped.

Table 2: Word correct rates of the baseline ASR system and closed caption in the specified acoustic conditions

<table>
<thead>
<tr>
<th>Condition</th>
<th>F0</th>
<th>F1</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed caption</td>
<td>77.1</td>
<td>74.8</td>
<td>66.8</td>
<td>79.3</td>
<td>74.9</td>
<td>77.1</td>
</tr>
<tr>
<td>Baseline ASR</td>
<td>87.4</td>
<td>64.0</td>
<td>81.5</td>
<td>77.2</td>
<td>24.6</td>
<td>77.8</td>
</tr>
</tbody>
</table>

Lack of ability in dealing with out-of-vocabulary (OOV) is another weakness of ASR. Even though our lexicon used in content indexing covers 300,000 words, it is still observed that 43 OOV words in WNT and 51 in NN appear in our test corpus.

2.4. Language Model Adaptation

Since new words are often introduced in broadcast news, we should update our language model in a timely fashion in order to capture them. To maintain our recognition accuracy consistently, we start to adapt the existing baseline language model to capture them. To maintain our recognition accuracy consistently, we start to adapt the existing baseline language model probabilities $P_{adapt}(w_t|h_t)$ as shown in Eq. 1.

$$P_{adapt}(w_t|h_t) = \lambda_0 \cdot P_0(w_t|h_t) + \lambda_{new} \cdot P_{new}(w_t|h_t)$$

where $P_0$, $P_{new}$ are the probabilities of the baseline model and the additional materials respectively, $\lambda_0$, $\lambda_{new}$ are the weights of each of the probabilities, $h_t = w_{t-1}, ..., w_{t-1}$ is the previous word-history for $w_t$, [12].
3. Phonetic Alignment

In this paper, our goal is to generate more accurate time-aligned transcriptions for content indexing, while previous work focused on detection of audio segments of which closed caption and ASR outputs match for automatic model adaptation. Our implementation tries to find the best alignment of two imperfect sources, closed captions and ASR outputs using phonetic alignment.

To synchronize closed captions, $c_i$, and ASR system outputs, $t_j$, a word-level alignment based on dynamic programming method shown at Eq.(2) had been used in a previous indexing system [13].

$$D(i, j) = \min \left\{ \begin{array}{ll} D(i, j - 1) + 1 \\ D(i - 1, j) + 1 \\ D(i, j - 1) + d_w(c_i, t_j) \end{array} \right. \quad (2)$$

where

$$d_w(c_i, t_j) = \left\{ \begin{array}{ll} 0 & \text{if } c_i = t_j \\ 1 & \text{if } c_i \neq t_j \end{array} \right.$$

The word-level alignment method could deal with the mismatches between captions and ASR outputs caused by insertion, deletion, and one-to-one substitution errors. However, in the case where mismatches involve one-to-many or many-to-many substitutions as shown in Table 3, then it becomes difficult to correct ASR errors by closed caption alignment.

<table>
<thead>
<tr>
<th>Closed Caption</th>
<th>ASR errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iraq /ax r aa k/</td>
<td>a rock /ax/ /r aa k/</td>
</tr>
<tr>
<td>YouTube /y uw t uw b/</td>
<td>you to /y uw/ /t uw/</td>
</tr>
<tr>
<td>Iraqi officials offer /ax r a k iy/ /ax f ih sh ax l /z /ao f er/</td>
<td>we’re accusations off for /iw iy t/ /ae k y uw z ey sh ax n z /ao e l /er/</td>
</tr>
<tr>
<td>morbid statistic only /m ao r b ih d /s t ax t ih s t ih k/ /ow n 1 iy/</td>
<td>mortgage to stick all week /m ao r g ih jh/ /t ax/ /s t ih k/ /ao e l /w iy k/</td>
</tr>
</tbody>
</table>

Table 3: Examples of the mismatches between closed captions and ASR outputs

It is still observed that the sounds from captions and ASR outputs are similar even though words from both don’t match. For example, pronunciation transcriptions of ‘Iraq’ and ‘a rock’ are similar. In this paper, another layer of alignment is introduced on top of the existing word-level alignment to recover ASR errors caused by sound similarity.

The proposed method is described in the following Algorithm 1. First, frequently occurring words, i.e., stop words, are removed from the set of tokens to be aligned because they often generate false alignments especially when words in the closed captions are missed. At the first stage, the word-level alignment described above is performed.

From the word-level alignment result, the portion is identified where the words from closed captions, $c_{s+1}...j$, and ones from ASR, $w_{s+1}...j$, differ. The words between a pair of anchor words, $c_s = w_{s+1}$, in the caption and ASR outputs are respectively transformed into each sequence of phoneme symbols.

Word-to-phoneme transformation is done by mainly dictionary look-up, while OOV words in the closed captions are transformed into each sequence of phoneme symbols. At the first stage, the word-level alignment penalizes the whole at mismatches caused by insertion, deletion, and one-to-one substitution errors. However, in the case where mismatches involve one-to-many or many-to-many substitutions as shown in Table 3, then it becomes difficult to correct ASR errors by closed caption alignment.

Algorithm 1 Phonetic alignment

Given caption, $C$, and ASR words, $W$.

// remove stop words, $G$, from $C$ and $W$
$c_i \in C \cap G'$; $w_j \in W \cap G'$

// 1st stage: word-level alignment
$D_P-word(c_i, w_j)$
$S^*_w = \text{Backtracking}(B_P)$

while $0 \leq i < j \leq T$ do

if $(c_i = w_{s+1}) \land (c_{s+1} \neq w_{s+1})$ then

// search anchor word pair

while $(c_i \neq w_j) \land (t(j) - t(i) < Max)$ do

$j \leftarrow j + 1$

end while

// convert caption and ASR words into phonemes
$p_1...M = \bigoplus_{i=1}^{M+1} \text{Dict}(c_{s+1}) || \text{TTS}(c_{s+1})$
$q_1...N = \bigoplus_{i=1}^{N+1} \text{Dict}(w_{s+1}) || \text{TTS}(w_{s+1})$

// 2nd stage: phone-level alignment
$D_p PHONE(p_m, q_n)$
$S^*_c = \text{Backtracking}(B_P)$

end if

$i \leftarrow j$; $j \leftarrow j + 1$

end while

function $D_P-WORD(i, j)$

return $\min [D(i, j - 1) + D(i - 1, j) + D(i, j - 1) + d_w(c_i, w_j)]$

end function

function $D_P-PHONE(i, j)$

return $\min [D(i, j - 1) + D(i - 1, j) + D(i, j - 1) + d_p(p_i, q_j)]$

end function

4. Experimental Results

Using the corpus as described in Section 2.1, word correct rates are measured from the ASR outputs using the baseline language
model (LM) and the ones using the LM adapted with the closed caption in the two-week period including the target date. Word-level alignment and phonetic alignment are applied on both the baseline ASR outputs and the adapted ones.

The word correct rate in the closed captions in our corpus ranged from 74.1% to 79.3% in all the conditions except F3 as shown in Table 4. In many cases, the titles of news programs and the anchor persons’ names presented with the background music do not appear in the closed captions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>F0</th>
<th>F1</th>
<th>F3</th>
<th>F4</th>
<th>F5</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caption</td>
<td>77.1</td>
<td>74.8</td>
<td>66.8</td>
<td>79.3</td>
<td>74.9</td>
<td>77.1</td>
</tr>
<tr>
<td>Baseline ASR</td>
<td>87.4</td>
<td>64.0</td>
<td>81.5</td>
<td>77.2</td>
<td>24.6</td>
<td>77.5</td>
</tr>
<tr>
<td>Base + W-Align</td>
<td>91.3</td>
<td>81.7</td>
<td>90.1</td>
<td>88.3</td>
<td>73.3</td>
<td>88.0</td>
</tr>
<tr>
<td>Base + P-Align</td>
<td>92.3</td>
<td>82.1</td>
<td>90.7</td>
<td>88.7</td>
<td>73.3</td>
<td>88.6</td>
</tr>
<tr>
<td>LM Adaptation</td>
<td>89.8</td>
<td>68.8</td>
<td>87.9</td>
<td>83.6</td>
<td>34.9</td>
<td>82.7</td>
</tr>
<tr>
<td>Adapt + W-Align</td>
<td>91.3</td>
<td>82.4</td>
<td>91.1</td>
<td>89.1</td>
<td>72.8</td>
<td>88.6</td>
</tr>
<tr>
<td>Adapt + P-Align</td>
<td>92.3</td>
<td>82.8</td>
<td>91.6</td>
<td>89.7</td>
<td>72.3</td>
<td>89.2</td>
</tr>
</tbody>
</table>

Table 4: Word correct rates of the baseline ASR system and the ASR using adapted language models in the specified acoustic conditions

Our experimental result shows that LM adaptation improves ASR accuracy on average 5% but the performance of the ASR using the adapted LM still suffers when recognizing spontaneous speech or acoustically degraded speech.

The word correct rates using the proposed phonetic alignment method are around 11% higher in the baseline ASR and 6.5% higher in the ASR using adapted LM. Especially when the acoustic condition is degraded, the proposed method has the benefit of improving transcription quality by recovering the ASR mistakes using human-transcribed captions.

The word correct rates in the phonetic alignment are only 0.6% higher than the ones from word-level alignment but phonetic alignment provides better localization of ASR errors as shown in Fig. 3. For example, the word, ‘YouTube’ was misrecognized as ‘you to’ by ASR.

Figure 3: An example of phonetic alignment with closed captions and ASR system outputs

Phonetic alignment finds the similarities from both sides, and pinpoints the exact location in audio when the word-level alignment can’t find a match, and later inserts ‘YouTube’ in the place of ‘you’ only. However, the phonetic alignment method is not free from false alarms caused by typographical errors in the closed caption. For example, these pairs such as send → spend, age → page, and sitting → sting are phonetically similar and correctly-spelled, but mistakes by captioners.

5. Conclusion

In this paper, a phonetic alignment method is introduced to improve transcription quality by finding phonetic similarities between two imperfect sources, closed captions and ASR outputs. On our broadcast news corpus, the proposed method shows word correct rate improvements of on average 11% from the ASR output using the baseline language model and 6% from the one using the adapted LM. It is expected that the proposed method enables more accurate content indexing and provides the capability of automatic learning of OOV words.

6. Acknowledgements

The authors would like to thank Barbara B. Holister and her labeling group for their hard work on transcribing the test audio thoroughly.

7. References