Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 35 (2016) 116–133

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

ALISA: An automatic lightly supervised speech segmentation and alignment tool☆,☆☆

A. Stan[a,*], Y. Mamiya[b], J. Yamagishi[a,c], P. Bell[b], O. Watts[b], R.A.J. Clark[b], S. King[b]

[a] *Communications Department, Technical University of Cluj-Napoca, 26-28 George Baritiu St., Cluj-Napoca 400027, Romania*

[b] *The Centre for Speech Technology Research, University of Edinburgh, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom*

[c] *National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan*

Received 14 November 2014; received in revised form 12 May 2015; accepted 23 June 2015
Available online 3 July 2015

Abstract

This paper describes the ALISA tool, which implements a lightly supervised method for sentence-level alignment of speech with imperfect transcripts. Its intended use is to enable the creation of new speech corpora from a multitude of resources in a language-independent fashion, thus avoiding the need to record or transcribe speech data. The method is designed so that it requires minimum user intervention and expert knowledge, and it is able to align data in languages which employ alphabetic scripts. It comprises a GMM-based voice activity detector and a highly constrained grapheme-based speech aligner. The method is evaluated objectively against a gold standard segmentation and transcription, as well as subjectively through building and testing speech synthesis systems from the retrieved data. Results show that on average, 70% of the original data is correctly aligned, with a word error rate of less than 0.5%. In one case, subjective listening tests show a statistically significant preference for voices built on the gold transcript, but this is small and in other tests, no statistically significant differences between the systems built from the fully supervised training data and the one which uses the proposed method are found.

Keywords: Speech segmentation; Speech and text alignment; Grapheme acoustic models; Lightly supervised system; Imperfect transcripts

1. Introduction

Over the past decade, speech-enabled applications have progressed to the point where their presence in human–computer interfaces is almost ubiquitous. However, this is true only for the languages for which a sufficient degree of effort has been invested in creating purposely built tools and resources. Any speech-based application

☆ This paper has been recommended for acceptance by Koichi Shinoda.

☆☆ This paper is based on our previous work (Stan et al., 2012, 2013; Mamiya et al., 2013), and presents it into a more coherent and thorough manner. Additional results and discussions are presented for the following key aspects: introducing more relaxed confidence measure conditions; grapheme-level acoustic likelihood scores within the confidence measure; unsupervised state tying for the tri-grapheme models; evaluation for an additional speech resource in a different language; and re-evaluation of our results using subjective listening tests for two languages.

* Corresponding author. Tel.: +40 264202452.
E-mail address: Adriana.Stan@com.utcluj.ro (A. Stan).

http://dx.doi.org/10.1016/j.csl.2015.06.006
0885-2308/

requires a large amount of high-quality data and expert knowledge, which are time consuming and computationally expensive to collect.

In this paper we try to alleviate one of the major problems which occurs when migrating a speech-based solution either from one language to another, or from one set of resources to another: speech data preparation. In both speech recognition and speech synthesis, the performance of the resulting system is highly dependent on the quality and amount of training data. But the most widespread method for gathering speech resources nowadays is either by recording a voice talent in a studio or by manually transcribing existing recorded data. However, both of these methods are tedious and usually deter developers from expanding their language and/or speaker environment portfolio.

Accordingly, we turn our attention towards the theoretically unlimited supply of speech data available on the internet, of which the majority is recorded in professional or semi-professional environments – an essential requirement to begin with. Examples of such data include audiobooks, podcasts, video lectures, video blogs, company presentations, news bulletins, etc. These resources are even more appealing when accompanied by an approximate transcript, even though the precise synchronisation between text and audio is not generally available.

The automatic alignment of speech and text has long been studied, and in Section 2 we present some of the most prominent methods. However, our goal is slightly different than theirs, as we aim to rely on no previous knowledge or prepared data, and try to provide a solution for any language with an alphabetic writing system.[1] This is of course highly error prone, but we will show in the Results section that the errors are minimal and negligible both for speech recognition and speech synthesis tasks.

The paper is structured as follows: in Section 2 we describe the state-of-the-art for speech and text alignment methods. Section 3 gives a brief overview of the proposed method, its individual steps being expanded in Sections 4–8. Objective and subjective evaluations of the tool are presented in Sections 9 and 10 respectively. Section 11 provides discussion and concludes the paper.

2. Related work

Approaches to the task of speech and text alignment can be divided into two major categories: ones where accurate orthographic or phonetic transcripts are available, and ones where errors and omissions occur in the transcripts.[2]

Within the first category, the task is simply to determine a direct correspondence between the acoustic units and the text symbols, whether they be sentences, words, letters or phones. This type of method is based either on well-trained pre-existing acoustic and language models, or on dynamic time warping algorithms (Anguera et al., 2011). One of the challenges presented by this type of approach is the alignment of long audio segments, for which an acoustic model based Viterbi decoder would require a large amount of computational resources. In Goldman (2011) the authors propose a phonetic alignment for English and French, under Praat using good acoustic models, utterance segmentation, grapheme to phoneme conversion, and manual intervention. Cerisara et al. (2009) proposes a GUI for speech alignment with integrated user feedback for manual checking and tuning, again relying on pre-existing acoustic models and speech segmentation (Moreno et al., 1998) determines a set of anchors within the speech data, and uses a recursive, gradually restrictive language model to recognise the text between two consecutive anchors.

In approaches from the second category, as well as determining the boundaries of sentences or words, the presence of transcription errors must also be taken into account. Due to this fact, most of the proposed methods rely on very good acoustic and language models. For example, Braunschweiler et al. (2010) use speaker-independent acoustic models previously trained on over 150 h of speech data in conjunction with a large, smoothed language model biased towards the text being aligned. Bordel et al. (2012) use a phone-level acoustic decoder without any language or phonotactic model and then finds the best match within the phonetic transcripts. This approach was a result of the fact that the data to be aligned may contain a mixture of languages. Prahallad and Black (2011) present a back-tracking method for Viterbi-based forced alignment and uses good acoustic models. It tries to align the data with 125 words, but it does not detail how to deal with long misalignments of missing audio/text segments. Moreno and Alberti (2009) introduce the use of a factor automaton as a highly constrained language model trained on the transcripts. It first divides the

[1] Other types of writing system might be viable, but within this work we do not investigate them.

[2] Take for example a recording of a trial, for which the typist provides a stenotype. This will most commonly include word deletions, substitutions or insertions, and sometimes additional comments.