Audio Surveillance Tapping Transcription Quality Estimation Reliability

Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
Email: {clara.borrelli, paolo.bestagini, fabio.antonacci, augusto.sarti, stefano.tubaro}@polimi.it

***Abstract*—The abstract goes here.**

I. Introduction

Thanks to the recent advances in technology, audio communication systems are becoming more and more pervasive and diverse. In fact, nowadays a conversation can happen not only face-to-face but also on a variety of channels, like phone calls, VoIP or voice messages on instant messaging platforms.

At the same time the ability of monitoring environmental, digital or phone communications has become an urgent necessity for national security issues. These technologies can effectively assist law enforcement agencies in foiling terrorist attacks or revealing harmful intents. The aforementioned heterogeneity of the communication channels has produced in this sense both advantages and drawbacks.

On one side, information gathering has become easier and ubiquitous. Regarding environmental monitoring, audio surveillance devices are getting cheaper and easier to deploy. Subsequently, the number of control spots can be further increased, ensuring higher coverage, for example, in large public spaces. Concerning digital or phone wire taps, *why is easier to collect data? because we have so many communication channels?*

On the other side, a blind and massive data collection can results in huge databases whose manual inspection might be infeasible. Tools like automatic transcription agents would be of great help for both investigative and surveillance purposes. However, given the diversity of context in which data is collected,the audio excerpts can be affected by several types of noise which can degrade the quality of the records and compromise the intelligibility of a possible relevant conversation.

For these reasons it is needed to develop automatic and intelligent methods that allow to speed up the analysis of these huge corpora but at the same time take in account the variety of the involved devices and environments. Depending on the characteristics of the collected audio excerpts, these systems should be able to extract relevant information while guaranteeing control to the human user.

In this work we try to meet these needs by proposing a framework able to evaluate the quality of noisy speech recordings. Given the specified context, this evaluation is based on the estimation of the likelihood of obtaining reliable transcript using a generic automatic speech-to-text engine.

Our method can be employed in two different applicative contexts. The first one is for *not sure what is the meaning of the bullet point. We want to propose an index to be exploited for tuning automatic transcription agents? or we want to compute the intelligibility index on large databases to evaluate the quality of a specific dataset?*. The second one is, as already mentioned, to provide a useful tool in managing data acquired by audio surveillance systems. The model provides a quality feedback to investigators, that will be able to discriminate reliable from non reliable automatically extracted transcriptions.

The main rationale is to use a data-driven approach. Our system is composed of two main blocks. In the first one, a suitable set of features is extracted from the original audio signal. This features provide a numerical description of the fundamental characteristics of the signal analysed. Then, both a regressor and a classificator are designed and trained to predict the quality level (or a discrete label) of noisy speech audio signal.

The implemented model has been trained and tested on a large dataset of transcribed speech signal augmented with several type of noises. The ground truth has been obtained starting from the variation between the original transcription and the transcription of the noisy signal. The obtained results have been evaluated using standard metrics and show the effectiveness of the proposed model.

II. Problem statement and background

In this section we first present a formal problem statement, specifying all the elements involved in the framework proposed. Secondly, we present a short review of the concept of intelligibility at the state of the art.

A. Problem formulation

An audio excerpt $x(t)$ composed by speech in a noisy environment can be written as

$$x(t) = s(t) + n(t)$$

, where $s(t)$ is the speech signal while $n(t)$ is the noise signal. Assume a speech-to-text model $STT(\cdot)$ is selected and that, given as input the audio signal, is able to recover the correspondent transcription. ... we want to estimate how much

WIFS'2019, December, 9-12, 2019, Delft, Netherlands. XXX-X-XXXX-XXXX-X/XX/$XX.00

it is useful for an analyst. This is done by estimating a score indicating how likely is it possible to correctly understand the speech.

B. Background

Ci sono tanti metodi in tante aree che si occupano di stima dell'intelligibilita'. Citiamono un po' e facciamo capire che sono tipicamente cuciti su scenari applicativi per noi non rilevanti (e.g., acustica di stanza, transmissioni telefoniche, etc.).

III. Method

Cappello introduttivo in cui si spiega che il framework e' composto da tre elementi

- La scelta di una misura di qualita' oggettiva (Jaccard, Coseno su vettori di parole)
- L'estrazione di descrittori (MFCC, etc.)
- Classificazione / regressione

A. Objective quality measure

B. Feature extraction

C. Decision taking

IV. Results

In this section bla bla ...

A. Dataset

Da dove siamo partiti, tipi di rumore, SNR, trascrittore.

B. Setup

Tipi di classificatore / regressore usati. Divisione train, validation e test (o forse in dataset).

C. Numerical analysis

risultati di classificazione / regressione con metodi diversi, e feature diverse.

V. Conclusions

It worked!

Acknowledgment

This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-16-2-0173. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

References

[1] A. Piva, "An overview on image forensics," *ISRN Signal Processing*, 2013.