The Multi-level Approach to Speech Corpora Annotation for Automatic Speech Recognition

Igor Glavatskih, Tatyana Platonova(✉), Valeria Rogozhina, Anna Shirokova,
Anna Smolina, Mikhail Kotov, Anna Ovsyannikova, Sergey Repalov,
and Mikhail Zulkarneev

Stel - Computer Systems Ltd, Moscow, Russia
{glavatskih_ia,platonova_ts,rogozina_vs,smolina_aa,
anna_a,kotov,anna}@stel.ru, repalov@gmail.com, zulkarneev@mail.ru
http://www.stel.ru

Abstract. In the paper the multi-level approach to audio files annotation is briefly summarized. The emphasis is mainly placed on the development of annotation rules. Firstly, some general requirements are outlined and more specific markers are listed, which may or may not be included in a particular rule set depending on the given practical task. Then software tools used for creating annotations and its spell-checking are described, and an example of a database created on the basis of the multi-level approach to annotation is given. Lastly, the application of tag sorting in ASR training and testing is discussed.

Keywords: Speech annotation · Speech corpus · Speech data · Speech recognition · Annotation guidelines · Orthographic transcript

1 Introduction

Speech corpora creation plays an essential role in language processing and is vastly used in practical tasks of speech technology such as automatic speech recognition. The quality of these systems' output strongly corresponds with the quality of the database used for their training.

As highlighted in [1], there are several steps in speech corpus development, and the most important of them are speech data collection and annotation. The techniques and methods used here mostly depend on the purpose a database is intended for. Thus, data can be collected online via mobile devices and web applications as in [2], or recorded in a studio as in [1,3]. Speech itself can be either read or spontaneous. In case of read speech balanced phonetically rich sentences of optimal length, produced either manually [1] or automatically [2,4], can be used. This method simplifies the process of annotation to some degree, since the real utterances are usually not very different from the text. However, it must be pointed out that although it is helpful for some problems' solutions, the ASR systems' performance on spontaneous speech data would not be of a high quality in this case.

A. Ronzhin et al. (Eds.): SPECOM 2015, LNAI 9319, pp. 438–445, 2015.
DOI: 10.1007/978-3-319-23132-7_54

At the same time, spontaneous speech has recently become one of the speech technology's main focuses due to the need of enabling natural human-machine communication. State-of-the-art ASR tasks (such as voice web search, voice dictation applications, etc.) are not generally performed on high-quality noise-free signals produced by qualified speakers. So for better output the training material should match the type of speech and the environment the speech recognizer would run in. Therefore, speech corpora annotation method should provide a means to reflect not only linguistic content, but also some extralinguistic information within the utterance.

In consideration of the foregoing, a multi-level approach to annotation was designed, which covers several levels of information about speech data: broad orthographic transcription (word sequence actually pronounced and speech/non-speech events within utterances), the information about the signal as a whole and the recording environment, the speaker information. This method is used to collect large corpora of spontaneous speech on different languages for developing multilingual ASR [5].

It the present paper characteristics of this approach are discussed in details. Furthermore, the program tool used in the annotation process is described and an example of a database created is given.

2 Multi-level Method of Speech Material Annotation

2.1 General Requirements

The design of annotation rules depends on several issues. First of all, it is the audio data submitted for annotation. For instance, if the audio file is long, it is reasonable to mark out the speech fragments for further annotation. Besides, if experts are dealing with noisy signals, the preliminary automatic processing should be recommended to exclude signals not valid for any annotation.

The language of annotation should be taken into account as well (especially in the case of adapting existing rules to a new language), since each language has its own peculiarities and phonetic phenomena that do not occur in other languages and are worth marking (e.g. accents, umlauts and other diacritical signs; apostrophes).

All annotations are created with a view of a particular practical goal, which determines specific features of the rule set. However, there are some general guidelines that are to be followed irrespective of the annotation task (e.g. the speech material that is to be annotated):

- All words are written correctly in the context of the language's orthographic system (for some tasks an exception is made for regional variants that are spelled as pronounced in those cases and have special marks);
- The orthographic transcriptions are to be comprehensive: all the speech fragments and elements should be transcribed successively, all speech and non-speech events should be marked in accordance with the developed rules;

- All words are to be written in lower case in order to speed up the annotation process (the exceptions include spelled letters and abbreviations that are written in capitals);
- No punctuation marks or symbols are allowed; hyphens within words should be substituted by spaces;
- Numerals, money amounts and measures should be written down as words;
- Meaningful fillers (e.g. "well", "say") are to be transcribed;
- The use of contracted constrictions is not allowed (e.g. "october" and not "oct");
- The speech cut off by truncation of the wave form at the beginning or the end of the signal should be marked.

2.2 Specific Marks

Alongside with general rules our experts register a number of speech and non-speech events that could be detected in sound files. This requirement arises from the necessity of training an ASR system not to confuse them with words.

- In-speech events:
  - noise;
  - music;
  - applause;
  - IVR speech;
  - side speech (speech by any person not intended for the recognizer);
  - overlapped speech (speaker's speech that occurs at the same time as ome other person's speech);
  - fragmented or interrupted word;
  - target speaker's noise, breath sound, yawning, laughter, cough;
  - shouted phrases;
  - hesitation words.
- Speech anomalies:
  - unclear pronunciation;
  - wrongly stressed syllables;
  - reduced words;
  - mispronounced words;
  - words in the non-target language;
  - overlong sound quantity;
  - syllable-by-syllable pronunciation.

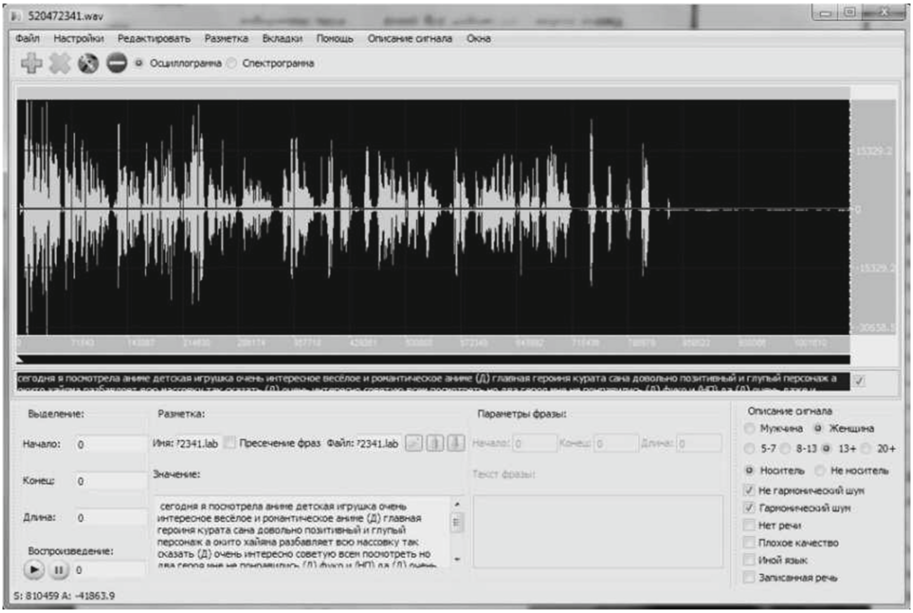
Furthermore, under our annotation rules the following signal and speaker features could be outlined (when occur):

- bad audio;
- harmonic background noise;
- inharmonic background noise;
- foreign (non-target) language;

- acoustic environment: speaker in studio, speech via telephone, prerecorded speech, voice-over translation, live commentary, film fragment, jingle;
- no speech;
- gender: male, female;
- age group: 5–8 years, 8–13 years, 13+ years, 20+ years;
- native/non-native.

2.3 Software Tools

In general the process of sound material annotation is a hard and time-consuming task since it is performed manually by human experts. Still some program tools could help to solve the problem. That is why Speech_Utility, a special data processing software, was designed. Speech_Utility simplifies the process to some degree, providing a means of creating an annotation, setting time labels for speech and tags, as well as marking some peculiar characteristics of signal (e.g. bad quality, harmonic/inharmonic noise, language, etc.) and speaker (e.g. male/female, age, native/non-native, etc.) by simply checking the boxes. The software interface (see Figs. 1, 2) allows the user to choose the type of image that is more appropriate for their task. Moreover, it has several additional plugins which contribute to the automation of preliminary processing. They automatically spot voice activity and mark its boundaries on the timescale; detect signals that are too noisy and exclude them from the material for annotation;



● Осциллограмма ○ Спектрограмма

сегодня я посмотрела аниме детская игрушка очень интересное весёлое и романтическое аниме (Д) главная героиня курата сана довольно позитивный и глупый персонаж а ☑

Начало: 0

Конец: 0

Длина: 0

Имя: ?2341.lab □ Пресечение фраз Файл: ?2341.lab

сегодня я посмотрела аниме детская игрушка очень интересное весёлое и романтическое аниме (Д) главная героиня курата сана довольно позитивный и глупый персонаж а окито хайяма разбавляет всю массовку так сказать (Д) очень интересно советую всем посмотреть но два героя мне не понравились (Д) фуко и (НП) ая (Д) очень

Начало: 0 Конец: 0 Длина: 0

Текст фразы:

Описание сигнала

○ Мужчина ● Женщина

○ 5-7 ○ 8-13 ● 13+ ○ 20+

● Носитель ○ Не носитель

☑ Не гармонический шум

☑ Гармонический шум

□ Нет речи

□ Плохое качество

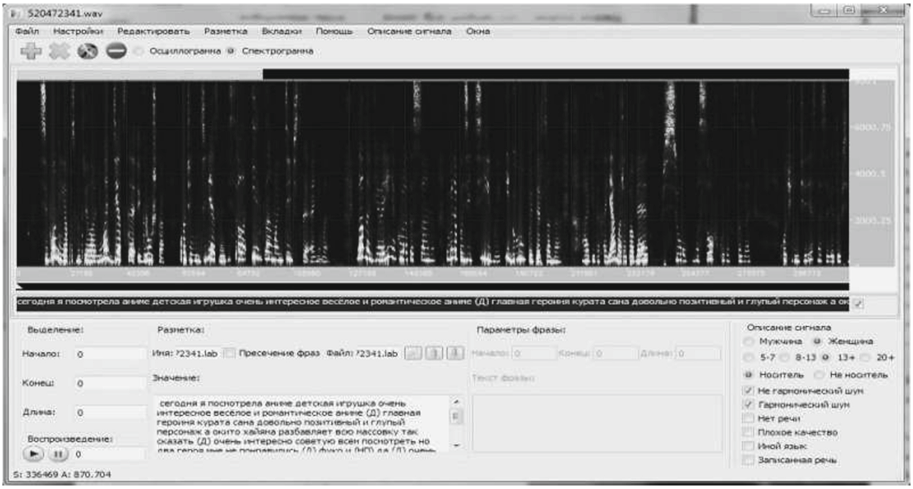
□ Иной язык

□ Записанная речь

0

S: 810459 A: -41863.9

Fig. 1. Speech_Utility software interface, oscillogram



сегодня я посмотрела аниме детская игрушка очень интересное весёлое и романтическое аниме (Д) главная героиня курата сана довольно позитивный и глупый персонаж а ок

Начало: 0

Конец: 0

Длина: 0

Воспроизведение: 0

Имя: ?2341.lab □ Пресечение фраз Файл: ?2341.lab

Значение:

сегодня я посмотрела аниме детская игрушка очень интересное весёлое и романтическое аниме (Д) главная героиня курата сана довольно позитивный и глупый персонаж а окито хайяма разбавляет всю массовку так сказать (Д) очень интересно советую всем посмотреть но два героя мне не понравились (Д) фуко и (НП) да (Д) очень

Параметры фразы: Начало: 0 Конец: 0 Длина: 0

Описание сигнала
○ Мужчина ● Женщина
○ 5-7 ○ 8-13 ● 13+ ○ 20+
● Носитель ○ Не носитель
☑ Не гармонический шум
☑ Гармонический шум
□ Нет речи
□ Плохое качество
□ Иной язык
□ Записанная речь

S: 336469 A: 870.704

Fig. 2. Speech_Utility software interface, spectrogram

generate phonetic transcriptions, setting time labels for each phoneme, reducing the experts' task to manual correction. Thus, Speech_Utility gives an opportunity to create a structured database that consists of audio signals, corresponding orthographic transcriptions and information about signals and speakers.

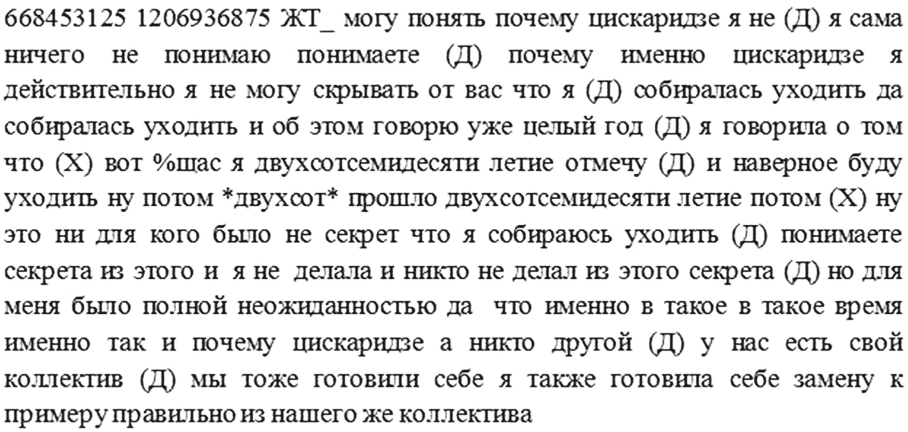
It must be mentioned that to ensure the consistency in spelling of all experts, a spell-checker is usually applied to every annotated word. In our work Yandex.Speller is used (for more information about this tool see [6]).

2.4 Example

As it has been already mentioned, annotation rules are typically framed for the purpose of corpora build-up. At the moment our research group has several ongoing projects related to this task. One of them is the speech database of broadcast speech. It is being created for the purposes of training and adapting acoustic models for automatic recognition of speech on different languages (Russian, English, German) obtained from various mass media sources. The Russian material is taken from web pages of different stations' TV and radio programs (e.g. from Euronews, Business FM, NTV, Russia Today, etc.). It should be mentioned that since the material is close to spontaneous speech (the sources of the speech data include broadcasting studio speech, on-the-spot reports, interviews, etc.), it contains a considerable amount of speech and non-speech events produced by a wide range of speakers. To gain more accurate recognition results the training database is to reflect them. The already annotated audio materials for the database of Russian broadcast speech include 75 h. Although the size of a database influences the quality of an ASR system output, it is correct speech data pre-processing that counts the most. For that reason all annotations are created manually by our experts on the basis of a particular rule set. For this task the Speech-Utility program is used and the main rules are:

- the following speaker categories are marked: male, female, child;
- the following types of acoustic environment are specified: speaker in studio, speech via telephone, prerecorded speech, voice-over translation, live commentary, film fragment;
- the following non-speech events are marked: music, inharmonic noise, target speaker's noise, breathing, coughing and laughter;
- the following speech events are marked: speech in a foreign language, unclear pronunciation, hesitation;
- the following speech events are marked and annotated: overlapped speech, shouted phrases, incorrect pronunciation, reduced words.

The resulting speech database includes annotation files containing orthographic transcriptions and information about the signal and speaker (see Fig. 3).



668453125 1206936875 ЖТ_ могу понять почему цискаридзе я не (Д) я сама ничего не понимаю понимаете (Д) почему именно цискаридзе я действительно я не могу скрывать от вас что я (Д) собиралась уходить да собиралась уходить и об этом говорю уже целый год (Д) я говорила о том что (Х) вот %щас я двухсотсемидесяти летие отмечу (Д) и наверное буду уходить ну потом *двухсот* прошло двухсотсемидесяти летие потом (Х) ну это ни для кого было не секрет что я собираюсь уходить (Д) понимаете секрета из этого и я не делала и никто не делал из этого секрета (Д) но для меня было полной неожиданностью да что именно в такое в такое время именно так и почему цискаридзе а никто другой (Д) у нас есть свой коллектив (Д) мы тоже готовили себе я также готовила себе замену к примеру правильно из нашего же коллектива

Fig. 3. An example of an annotation file containing an orthographic transcription

The acoustic models that were trained on the database of mass media language are incorporated in the Audioprotocol software system [8].

3 Tag Sorting for ASR

In respect to ASR the process of sound files annotation is mainly applied to build databases for training and testing. The relevancy of tags in annotations hugely varies depending on task requirements. Tag sorting of annotated data is used to form either test set or training set.

Test sets are framed with regard to the following points:

- All the words should be spelled orthographically correctly, regardless of the way they were pronounced; this condition is important since ASR systems are expected to output valid transcriptions.
- Annotation marks are used to exclude particular signals from the test set and to create a test sample that meets certain requirements; that is why tags that describe the whole signal (e.g. speaker's gender, age, type of acoustic environment or back-ground noise, etc.) are more appropriate for the task.

Training sets should fit specific requirements as well:

- Speech and non-speech events should be marked, since training sets are usually grouped on the basis of these tags.
- Signal features that make it invalid for ASR processing should be marked (e.g. speech overlapped by music or by speech, etc.).
- Wrongly stressed syllables should be marked, so that rule-based letter-to-sound converter [7] could build the right phonetic transcription (since stressed syllables do not follow the same pattern in pronunciation as non-stressed ones they should be treated separately).
- Transcriptions with incorrect spelling that imitates speaker's pronunciation are used for acoustic models' training in order to make ASR system understand and identify different regional variants of the same word.
- For the purpose of preventing statistic distortion only orthographically correct transcriptions should be applied to language modeling.

4 Conclusion

The present paper describes basic aspects of our approach to speech annotation, applied in ASR systems development. The method could be referred to as multi-level, since it implies covering several levels of information concerning speech data. Thus, the annotations include text transcriptions along with special markers for various speech/non-speech events as well as the information about signal and speaker features.

It should be pointed out that the developed approach meets two fundamental requirements that arise in relation to speech annotating. On the one hand, it enables creating adequate annotations that represent all the basic parameters relevant for ASR systems; on the other hand, the approach facilitates and speeds up the process of annotation, making it more convenient for experts.

Acknowledgments. Research is conducted by Stel - Computer Systems Ltd. with support of the Ministry of Education and Science of the Russian Federation (Contract #14.579.21.0058) Unique ID for Applied Scientific Research (project) RFMEFI57914X0058. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

References

1. Bogdanov, D.S., Brukhtii, A.V., Krivnova, O.F., Podrabinovich, A.Ya., Strokin, G.S.: The technology of speech databases formation. Collected papers of system

Analysis Institute of RAS, pp. 238–259. Editorial URSS, Moscow (2003–2004) (in Russian)

2. Lane, I., Wailbel, A.: Tools of collecting speech corpora via mechanishanical-truk. In: 11th NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Truck, California, pp. 184–187 (2010)
3. Matoušek, J., Romportl, J.: Recording and annotation of speech corpus for Czech unit selection speech synthesis. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 326–333. Springer, Heidelberg (2007)
4. Anumanchipalli, G., Chitturi, R., Joshi, S., Kumar, R., Singh, S.P., Sitaram, R.N.V., Kishore, S.P.: Development of Indian language speech databases for large vocabulary speech recognition systems. In: 10th SPECOM International Conference on Speech and Computer, Patras, pp. 245–254 (2005)
5. Zulkarneev, M., Grigoryan, R., Shamraev, N.: Acoustic modeling with deep belief networks for Russian speech recognition. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM 2013. LNCS, vol. 8113, pp. 17–24. Springer, Heidelberg (2013)
6. YandexSpeller. https://tech.yandex.ru/speller/
7. Krivnova, O.F., Zakharov, L.M., Strokin, G.S.: Automatic transcriber of russian texts: problems, structure and application. In: 6th SPECOM International Conference on Speech and Computer, Moscow, pp. 408–409 (2001)
8. Audiprotocol. http://speech.stel.ru:8080 (online access: login guest, password 1)