Arabic dialect identification based on probabilistic-phonetic modeling

Naim TERBEH✉, Mounir ZRIGUI

LaTICE Laboratory
Faculty of sciences of Monastir, Monastir 5000-Tunisia
naim.terbeh@gmail.com, terbehnaim1987@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract. The identification of Arabic dialects is considered to be the first pre-processing component for any natural language processing problem. This task is useful for automatic translation, information retrieval, opinion mining and sentiment analysis. In this purpose, we propose a statistical approach based on the phonetic modeling to identify the correspondent Arabic dialect for each input acoustic signal. The main idea consists first, and for each dialect, in calculating a referenced phonetic model. Second, for every input audio signal, we calculate an appropriate phonetic model. Third, we compare this latter to all referenced Arabic dialect models. Finally, we associate the input acoustic signal to the dialect where the referenced phonetic model minimizes the cosine similarity. The obtained results are satisfactory. Indeed, based on 117 audio sequences, we attain a classification rate of 93%. Supporting the achieved results and the coverage of most of Arabic dialects, this study can be a reference for future work addressing dialectical speech processing applications.

Keywords: Arabic dialects, probabilistic-phonetic model, dialect identification, cosine similarity.

1 Introduction

Human-machine communication is in full progress, thus facilitating the accessibility to information and its treatments by introducing new faster methods to access information like voice commands. Nonetheless, the dialectal variability can prevent much understanding of the vocal command. In order to assist interactive systems in the comprehension of transmitted messages, several studies addressing the Arabic dialect identification have been established.

The Arabic dialect identification has become the central task for most applications of Arabic speech processing, such as machine translation, speech recognition or social media analysis. In accordance with Zaidan et al. [1], dialect identification can be seen as an application of language identification applied to a group of closely related languages.

The literature contains recent work that has proposed statistical approaches for Arabic dialect identification. However, current methods are often based on linguistic resources (corpora, lexicon, dictionaries), which does not always exist, especially for Maghrebi Arabic. For this reason, we propose a combination between linguistic and numerical methods to identify the dialectal origin for each input audio signal. The sample of dialects covers Tunisian, Algerian, Moroccan, Syrian, Palestinian and Egyptian.

This paper is structured as follows. The different dialects of spoken Arabic language are briefly described in section 2. In section 3, we expose some work from the literature addressing the Arabic dialect identification. The proposed methodology is detailed in section 4. The details about the experiments are described in section 5. The concluding remarks and future work are mentioned in section 6.

2 Dialectical variability

The speakers of the Arabic language use in their daily discourses various dialects that can be considered as alternatives of the Modern Standard Arabic (MSA). Tunisian, Algerian and Moroccan dialects share several phonological traits among themselves thanks to their common history. Their lexicon contains several pronunciations inherited from other languages like Berber, French, Turkish, Italian and Spanish. Also, Syrian, Palestinian and Egyptian dialects share a lot of phonological features. In the following sub-sections, we briefly describe these dialects.

2.1 Tunisian dialect

Similar to other Maghrebi dialects, Tunisian vocabulary is generally Arabic, with some Berber words. However, it is morphologically and phonologically different from the MSA. The Tunisian dialect is very agglutinative: Speakers use often very few words where just one expresses a whole sentence. It differs from the MSA especially in its negation form where the markers are always agglutinated to other words as affixes or suffixes. Moreover, in the Tunisian dialect, several Arabic words are used with significant changes in their stem formation.

2.2 Algerian dialect

The Algerian dialect is an informal spoken language, not used in official speech. Its vocabulary is roughly similar throughout Algeria. Nevertheless, in the east of the country, the dialect is closer to the Tunisian one whereas in the west it is closer to the Moroccan one. Most of the words of Algerian dialect come from the MSA [2], but there is a significant variation in vocalization in most cases, and some omission of some sounds in other cases. Contrary to the MSA, few sounds are not used in Algerian discourses like ظ and ذ, where most of the time they are respectively pronounced as ض and د. Furthermore, the Algerian dialect uses some non-Arabic sounds like ڤ and پ.

2.3 Moroccan dialect

The Moroccan dialect or the Moroccan Darija is a member of the Maghrebi Arabic language continuum spoken in Morocco. It is mutually intelligible to some extent with the Algerian dialect and to a lesser extent with Tunisian one. It has been significantly influenced by other vocabulary like Berber, Latin, French and Spanish.

2.4 Egyptian dialect

The Egyptian Arabic is a North African dialect of the Arabic language, which is a branch of the Afro-Asiatic language. It originated in the Nile in Lower Egypt around the capital Cairo. Egyptian Arabic evolved from the Quranic Arabic, which was brought to Egypt during the seventh-century Muslim conquest that aimed to spread the Islamic faith among the Egyptians. The Egyptian dialect was very highly influenced by the Coptic language which was the native language of the Egyptians prior to the Arab conquest [3], and later it was significantly influenced by other languages such as French, Italian, Turkish and English.

2.5 Syrian and Palestinian dialects

The Syrian and Palestinian dialects are part of Levantine Spoken Arabic which covers also Lebanese and Jordanian dialects. Phonologically, structurally and lexically, we can mention several common features between Levantine Arabic and other varieties of Arabic. On the other hand, there are significant differences among Levantine dialects based on geographical areas and urban/rural division. The Syrian dialect is highly influenced by the Syrian language, a Semitic language of the Middle East, which belongs to the Aramaean language group and contains a large vocabulary inherited from Turkish and French languages. The Palestinian dialect presents phonetically slightly different compared to north Levantine dialects. It can follow two main varieties: urban and countryside. It can be also classified geographically into north and south.

3 State of the art

The literature has presented multiple studies addressing dialect identification. In such work, researchers have tried to develop new platforms whose goal has been to associate the adequate dialect to each input acoustic signal. We can mention the following:

- In the objective of identifying Arabic and Chinese dialects, Zhang et al. recommended a novel study based on the frequency of common n-grams. The main idea is to classify the input acoustic signal in the class which maximizes the number of n-grams compared to a reference acoustic base. It is the target of the work in [5]. Compared to the existing systems, the obtained results showed a strong correlation between dialect-salience and the frequency of occurrences in n-grams.

- To identify Jordanian and Egyptian dialects, Al-Ayyoub et al. invented in [6] a novel methodology based on the combination between different numerical and linguistic audio techniques. Through this study, the authors put forward a new solution to the problem of dialect identification and determined as well the combination of features/classifiers that would generate the best results. Based on a large corpus of Jordanian and Egyptian dialects, the suggested system showed a good performance.
- In [7], Guellil et al. proposed an unsupervised approach to identify the Algerian dialect within social media. In order to do so, the authors used a large Algerian dialectal lexicon. The proposition was based on the improved Levenshtein distance [11, 12]. Supporting a corpus of 100 messages that were collected using the Facebook API, the authors obtained an identification rate exceeding 60%.
- Based on the naive Bayesian algorithm and a transfer system, Hamada et al. put forward in [4] a novel study addressing the identification of the Egyptian dialect from text messages that circulated on social networks and generated the corresponding MSA representation. Using 3,000 words presenting the Egyptian dialect, the authors attained an identification rate which exceeded 92%.

In spite of the richness of the literature with studies addressing Arabic dialect identification, the intra-dialect variability presents a motivation to propose new robust features and new measures of similarity. Our contribution consists, for each input acoustic signal, in introducing a new methodology whose target is to calculate its similarity compared to referenced dialectal bases.

4 Methodology

The goal of this paper is to assign each input acoustic signal to an adequate Arabic dialect. The main idea consists in comparing between the phonetic model representing the input acoustic signal and the referenced models of different Arabic dialects. In accordance with the results of this comparison, we assign the input acoustic signal to the class which minimizes the cosine similarity. Figure 1 describes the operation of the proposed system:

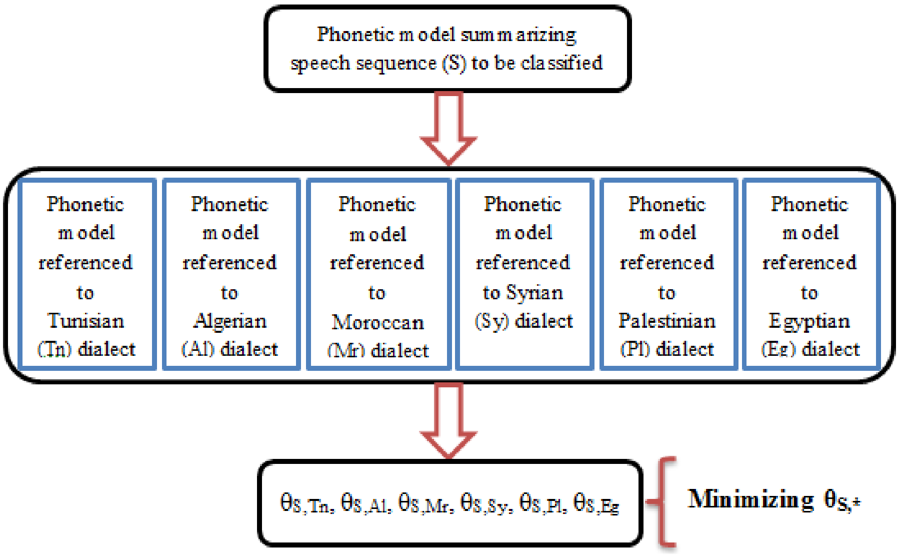
Phonetic model summarizing speech sequence (S) to be classified
Phonetic model referenced to Tunisian (Tn) dialect
Phonetic model referenced to Algerian (Al) dialect
Phonetic model referenced to Moroccan (Mr) dialect
Phonetic model referenced to Syrian (Sy) dialect
Phonetic model referenced to Palestinian (Pl) dialect
Phonetic model referenced to Egyptian (Eg) dialect
$\theta_{S,Tn}$, $\theta_{S,Al}$, $\theta_{S,Mr}$, $\theta_{S,Sy}$, $\theta_{S,Pl}$, $\theta_{S,Eg}$
Minimizing $\theta_{S,*}$

Figure 1. General form of proposed methodology

4.1 Acoustic modeling

The generation of phonetic models referenced to Arabic dialects requires an acoustic model trained to different spoken Arabic varieties and a large base of dialectical speeches. The speech base must be recorded by native speakers and cover all dialectical variabilities. The following table (Table 1) summarizes the corpora used to train our acoustic model.

Table 1. Summary of base of speeches used to train the acoustic model

Objective		Speech bases and sizes
Training an acoustic model for dialectical Arabic language	388 minutes	77 minutes of dialectical Tunisian speeches
		57 minutes of dialectical Algerian speeches
		66 minutes of dialectical Moroccan speeches
		62 minutes of dialectical Syrian speeches
		70 minutes of dialectical Palestinian speeches
		56 minutes of dialectical Egyptian speeches

4.2 Forced alignment

Supporting the Sphinx_align tool, we make the forced alignment procedure that allows generating for each speech signal the suitable phonetic transcription. For this treatment, it is sufficient to give this tool the paths to the necessary data, which are the voiced signals in the MFCC format, the suitable phonetic transcriptions, the pronunciation dictionary and the acoustic model.

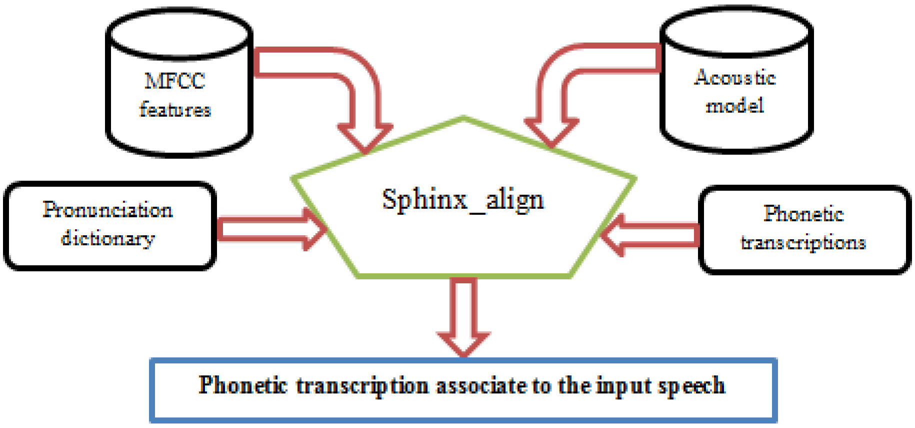

Figure 2. Sphinx_align procedure

4.3 Phonetic similarity and classification

The obtained phonetic transcription (previous sub-section) will be decomposed by bi-phonemes and the probability of occurrence for each bi-phoneme will be calculated. The vector arranging all these probabilities forms the phonetic model referenced to the dialect covered by the input speeches. Table 2 gives details about the base of speeches used to generate dialectic phonetic models.

Table 2. Summary of base of speeches used to calculate dialectical phonetic models

Objective		Speech bases and sizes
Calculating dialectical phonetic models	413 minutes	88 minutes of dialectical Tunisian speeches
		67 minutes of dialectical Algerian speeches
		78 minutes of dialectical Moroccan speeches
		51 minutes of dialectical Syrian speeches
		58 minutes of dialectical Palestinian speeches
		71 minutes of dialectical Egyptian speeches

We must guarantee a correlation between all phonetic models in the presentation and in the order of bi-phonemes. For example, the Algerian dialect includes the phonemes “ڤ” and “پ”, which is not the case for other Arabic dialects, so all models must comprise bi-phonemes containing these phonemes in the same order compared to the phonetic model referenced to the Algerian dialect.

Figure 3 illustrates an extract from the phonetic model referenced to the Tunisian dialect.

$$\begin{bmatrix} H\_D \\ H\_DH \\ H\_R \\ H\_Z \\ H\_S \\ H\_SH \\ H\_SS \\ H\_DD \\ H\_TT \\ H\_DH2 \\ H\_AI \\ H\_GH \\ H\_F \\ H\_Q \\ H\_K \end{bmatrix} = \begin{bmatrix} \mathbf{0.0015423348\%} \\ \mathbf{0.0046362397\%} \\ \mathbf{9.346364E-4\%} \\ \mathbf{1.4776867E-5\%} \\ \mathbf{0.0\%} \\ \mathbf{6.8342975E-5\%} \\ \mathbf{0.0\%} \\ \mathbf{3.140083E-5\%} \\ \mathbf{5.5413225E-6\%} \\ \mathbf{0.0\%} \\ \mathbf{0.0\%} \\ \mathbf{0.0\%} \\ \mathbf{3.6942151E-6\%} \\ \mathbf{3.8789258E-5\%} \\ \mathbf{1.4222728E-4\%} \end{bmatrix}$$

Figure 3. An extract from Tunisian dialectical phonetic model

For each new speech (S) to be associated to a suitable dialect, we transform this acoustic signal to its phonetic form (phonetic model), following the same procedure in section 4.2. We calculate, thereafter, all angles $\theta_{S,i}$ (i=1, …, 6) which separate this model for each one of the dialectical sample studied in this paper (Tn: Tunisian, Al: Algerian, Mr: Moroccan, Sy: Syrian, Pl: Palestinian and Eg: Egyptian). Finally, the input speech will be assigned to the dialect that minimizes the angle of similarity.

For example, a speech sequence S is associated to an appropriate dialectical class as follows:

- We calculate $\Theta = \{\theta_{S,Tn}, \theta_{S,Al}, \theta_{S,Mr}, \theta_{S,Sy}, \theta_{S,Pl}, \theta_{S,Eg}\}$ is the set of all phonetic similarity distances between the input speech sequence and all other spoken Arabic varieties.
- We suppose Min(Θ) is the minimum of the set Θ.
- The dialect which verifies the Min(Θ) is the most adequate to the input speech sequence.

The calculation of the set Θ is based on the following scalar product formulas:

$V_1 . V_2 = \sum_{i=1}^{n} V_1[i] . V_2[i]$; V_1 and $V_2$are two vectors representing two different phonetic models (1)

$V_1 . V_2 = \|V_1\| . \|V_2\| . \cos(\alpha)$; α is the angle that separate between V_1 and V_2 (2)

Thus, we conclude:

$$\cos(\alpha) = {}^{V_1 . V_2}\!/_{\|V_1\| . \|V_2\|} \quad (3)$$

5 Tests and results

5.1 Test conditions

The test is done under the following conditions:

- We prepare an acoustic model trained to dialectical Arabic speeches. For this objective we utilize a voice corpus of 388 minutes of Arabic speech which covers Tunisian, Algerian, Moroccan, Syrian, Palestinian and Egyptian dialects (on *.wav format and in mono speaker mode).
- We calculate six phonetic models, one for each Arabic dialect covered by this study. We use, for this goal, an acoustic base containing 413 minutes of Arabic dialectical speeches (on *.wav format and in mono speaker mode).
- We test the performance of the suggested method based on 117 speech sequences recorded by native speakers, which covers all the studied Arabic spoken varieties in this paper. All records follow the *.wav format and the mono speaker mode. The repartition of the test base is as follows:

Table 3. Summary of base of speeches used to evaluate proposed methodology

Objective	Utilized test base	
Testing proposed methodology	117 sequences	26 sequences of dialectical Tunisian speeches
		21 sequences of dialectical Algerian speeches
		15 sequences of dialectical Moroccan speeches
		17 sequences of dialectical Syrian speeches
		22 sequences of dialectical Palestinian speeches
		16 sequences of dialectical Egyptian speeches

5.2 Experimental results

In this section, we are interested in measuring the similarity between each pair of Arabic dialects through phonetic models that reference different Arabic spoken varieties. For this purpose, we choose to use the cosine similarity [8, 9], a measure of correlation between documents. It quantifies the similarity between two phonetic models. The choice of this metric is justified by its performance guaranteed in the document comparison [10].

The following table (Table 4) illustrates the results of dialectical speech classification and describes the confusion rate inter-dialects.

Table 4. Classification results using proposed method

Test bases	Classification results					
	Tn	Al	Mr	Sy	Pl	Eg
Tn	92.30%	0%	7.69%	0%	0%	0%
Al	0%	95.23%	4.76%	0%	0%	0%
Mr	0%	6.66%	93.33%	0%	0%	0%
Sy	0%	0%	0%	94.11%	5.88%	0%
Pl	0%	0%	0%	9.09%	90.90%	0%
Eg	0%	0%	0%	0%	6.25%	93.75%

Figure 4 details the performance of our proposed approach for each Arabic dialect.

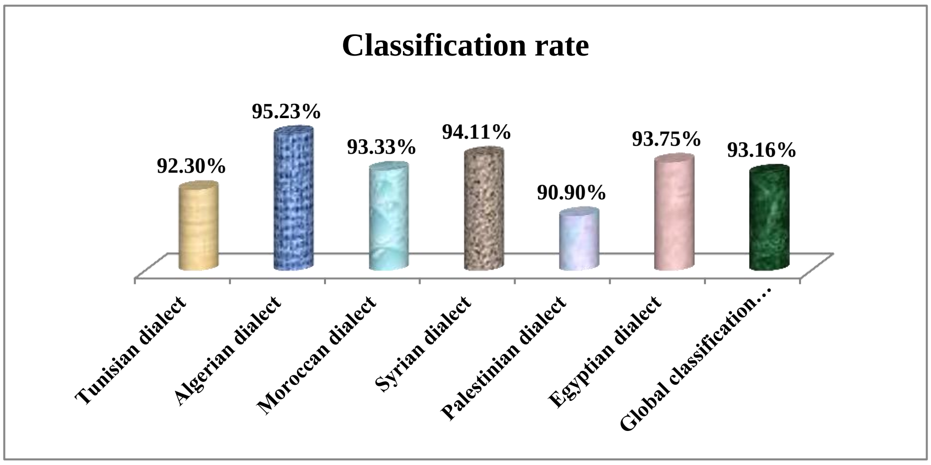

Figure 4. Classification rate for each Arabic dialect

5.3 Discussion

Table 4 draws some confusions between Arabic dialects. It is clearly shown that the highest confusion rates are those between Algerian and Moroccan and between Palestinian and Syrian dialects. This confusion is justified by the closeness between these pairs of dialects; e.g., Palestinian and Syrian dialects share significant vocabulary.

Some misclassification speech sequences can be justified by the shortness of the acoustic signal. Indeed, the dialectical speech sequence to be classified is short. Probably, the phonetic model does not cover all possible bi-phonemes, so the similarity with the referenced phonetic model will be falsified.

6 Conclusions and future work

To conclude, we can mention that the comparison between phonetic models presents a deciding factor to classify Arabic dialectical speeches. In this paper, we have put forward a probabilistic-phonetic methodology to assign each input acoustic signal to a suitable Arabic dialect. For this purpose, a corpus of 413 minutes of Arabic dialectical speeches has been prepared to calculate the phonetic models referring to the different spoken Arabic dialects. Another corpus containing 388 minutes of Arabic speeches covering six Arabic dialects has been recorded to train the acoustic model.

Based on 117 speech sequences, our proposed method has presented a high performance. Indeed, we have had 93% as a classification rate of Arabic dialects and we have extracted some confusion inter-dialects that confirm the closeness between these dialects.

To the best of our knowledge, this work presents the widest dialectical coverage. We are satisfied with the obtained results, and our suggested approach can present an important reference for work focalizing on the classification of dialectical speeches.

As future work, we can extend this study to elaborate a new platform whose goal is to transform an input acoustic signal from the dialectical form to its adequate one in MSA [13, 14].

Acknowledgements

First, I would like to express my deepest regards and thanks to all members of the scientific committee within the Computación y Sistemas journal. I extend also my advance thanks to my supervisors for their valuable advices and encouragement.

References

1. ZAIDAN O. F., CALLISON-BURCH C. (2011). The Arabic online commentary dataset: An annotated dataset of informal Arabic with high dialectal content. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT’11, p. 37–41, Stroudsburg, PA, USA: Association for Computational Linguistics.
2. Meftouh, K., Bouchemal, N., Smaili, K.: A Study of a Non-resourced Language: an Algerian Dialect. In: Third International Workshop on Spoken Languages Technologies for Under-resourced Languages. (2012) 125–132.
3. Nishio, Tetsuo. "Word order and word order change of wh-questions in Egyptian Arabic: The Coptic substratum reconsidered". Proceedings of the 2nd International Conference of L'Association Internationale pour la Dialectologie Arabe. Cambridge: University of Cambridge. 1996, pp. 171-179.
4. Salwa Hamada, Reham M. Marzouk: Developing a Transfer-Based System for Arabic Dialects Translation. Intelligent Natural Language Processing: Trends and Applications vol. 740 (2018) 121-138.
5. Qian Zhang ; Hynek Bořil ; John H. L. Hansen : Supervector pre-processing for PRSVM-based Chinese and Arabic dialect identification. IEEE International Conference on Acoustics, Speech and Signal Processing, 2013. ICASSP 2013.
6. Mahmoud Al-Ayyoub ; Marwan K. Rihani ; Nidal I. Dalgamoni ; Nawaf A. Abdulla : Spoken Arabic dialects identification: The case of Egyptian and Jordanian dialects. 5th International Conference on Information and Communication Systems (ICICS), 2014
7. Imène Guellil ; Faiçal Azouaou : Arabic Dialect Identification with an Unsupervised Learning (Based on a Lexicon). Application Case: ALGERIAN Dialect. IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES), 2016.

8. Nguyen, H. V., & Bai, L. (2010, November). Cosine similarity metric learning for face verification. In Asian Conference on Computer Vision (pp. 709-720). Springer Berlin Heidelberg.
9. Ye, J. (2011). Cosine similarity measures for intuitionistic fuzzy sets and their applications. Mathematical and Computer Modelling, 53(1), 91-97.
10. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526).
11. Ho, T., Oh, S. R., & Kim, H. (2017). A parallel approximate string matching under Levenshtein distance on graphics processing units using warp-shuffle operations. PloS one, 12(10).
12. Lee, T. (2017). LEVENSHTEIN DISTANCE-BASED REGULARITY MEASUREMENT OF CIRCADIAN RHYTHM PATTERNS. Journal of Theoretical & Applied Information Technology, 95(18).
13. Harrat, S., Meftouh, K., & Smaili, K. (2017, April). Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING).
14. Malmasi, S., & Zampieri, M. (2017). Arabic dialect identification using iVectors and ASR transcripts. VarDial 2017.