ВЕРИФИКАЦИЯ ЛИЧНОСТИ ПО ГОЛОСУ С ИСПОЛЬЗОВАНИЕМ ПРОСОДИЧЕСКИХ ПАРАМЕТРОВ

Пояганова М.С. (mpoyaganova@yandex.ru)

НИУ ВШЭ, Нижний Новгород, Россия

Voice authentication as a biometrics type has demonstrated its relevance in recent years. This paper is dedicated to the task of automatic speaker recognition. Particularly, we focus on speaker verification, considered in terms of binary classification task. In contrast to the majority of speaker recognition systems based on Mel-frequency cepstral coefficients (MFCCs), we investigate the significance of speaker recognition system based on the prosodic parameters. To analyze the system performance we apply several widely-used machine learning methods (Balanced Bagging, Random Forest, kNN). We also make the advantage of some oversampling techniques to deal with class imbalance problem, typical for speaker verification task. Experiments were run using speech samples of 50 speakers taken from the TIMIT corpus.

Key words: speech processing, speaker identification, speaker verification, prosody, prosodic features

Введение

Годы разработок в области создания систем, способных идентифицировать личность говорящего по голосу, привели к значимым результатам. Большинство таких систем опирается на акустические модели, которые используют кепстральные признаки, игнорируя информацию на более высоком, сверхсегментном уровне. Этот уровень включает в себя просодику, охватывая временные, тональные и динамические параметры в речи. Системы, учитывающие просодические признаки, работают более точно, обладают лучшей способностью к пониманию естественной речи. Однако, по причине недостаточного внимания к просодике, а также сложностей, связанных с репрезентацией и моделированием просодических признаков, применение данных параметров в идентификационных системах еще не доведено до необходимого уровня.

Таким образом, данная работа посвящена исследованию просодических признаков, их роли в системах автоматического распознавания речи, а также применению модели, способной верифицировать личность говорящего, основанной только на просодических параметрах.

Актуальность настоящего исследования формируют следующие аспекты:

1) Широкая область применения систем верификации личности по голосу.

Сегодня данные системы активно используются для управления доступом к персональным данным, например, к банковскому счету или личным документам, хранящимся в электронном архиве. Кроме того, идентификационные системы находят широкое применение в криминалистике и судебной экспертизе.

2) Необходимость детального изучения вопроса распознавания личности по голосу для совершенствования технологий автоматической верификации личности.

3) Потребность в систематизации знаний о применении просодических признаков в моделях распознавания говорящего.

Несмотря на большое число попыток встроить просодические признаки в идентификационные/верификационные модели, исследователями до сих пор не решено однозначно, какой способ репрезентации и моделирования просодики является наиболее точным и, следовательно, предпочтительным.

Просодические признаки

Многие зарубежные исследователи применяли просодические признаки для задачи автоматического распознавания личности по голосу. Большинство работ посвящено анализу качества работы систем, основанных на комбинации мел-частотных кепстральных коэффициентов и просодических признаков (E. Shriberg, M. Farru's). Так, согласно исследованию Sönmez, добавление просодических параметров к задаче верификации личности по голосу улучшает качество кепстральной системы, построенной на смеси Гауссовских распределений на 10% (Sönmez, 1998).

Практическая значимость данного исследования состоит в применении просодики как основного материала, содержащего биометрическую информацию о спикере. Кроме того, использованные техники классификации, насколько нам известно, являются ранее не описанными применительно к исследуемой задаче.

Материал исследования

Процесс сбора данных для исследования был разделён на три этапа:

1) Поиск корпуса звучащей речи;

2) Аннотация аудиозаписей;

3) Автоматическое извлечение просодических признаков.

Для данного исследования был выбран корпус английской речи TIMIT (Garofolo, 1993). Данный корпус представляет собой набор данных, предназначенных для акустико-фонетических исследований речи, а также для разработки и оценки систем автоматического распознавания речи. Корпус TIMIT содержит аудиозаписи носителей восьми основных диалектов американского английского языка. Общее число дикторов в корпусе составляет 630 человек, каждому из которых соответствует 10 прочитанных ими фонетически богатых предложений, каждое хранящееся в виде отдельного аудиофайла в формате .wav. Применительно к данному исследованию, корпус TIMIT не использовался полностью, были отобраны по 5 аудиозаписей для 50 дикторов. (25 лиц мужского пола и 25 лиц женского пола для сбалансированности выборки). Каждая из 250 выбранных аудиозаписей имела абсолютную длительность от 2 до 7 секунд.

После отбора аудиозаписей из речевого корпуса осуществлялась их аннотация. Инструментом, применявшимся на этом этапе исследования стала программа Praat (Boersma P., Weenink D). Данная программа позволяет проводить анализ звучащей речи, синтез речи по артикуляционным, фонетическим и акустическим характеристикам речи, аннотирование и сегментацию. Аннотация аудиозаписей осуществлялась по слогам. Такой тип аннотирования был выбран как наиболее оптимальный в данной задаче – на слоговом уровне фиксируются необходимые просодические параметры (Farru's, 2008).

Предобработка данных

В рамках предобработки данных была проведена нормализация признаков. Так, с опорой на исследования распознавания личности по голосу с применением просодики (Farru's, 2008), было решено логарифмировать значения признаков. Выбросы и шумовые значения отсутствовали в наборе данных, поскольку он был сформирован вручную, с учетом всех ограничений. Однако в наборе присутствовали пропущенные значения по признакам f min, f max, f mean: 42 наблюдения из 2845. Пропущенные значения в тональных параметрах появлялись по причине недостаточно длительного периода вокализации в некоторых слогах, то есть, мелодика речи просто не отражалась на данных звуковых участках. Удаление объектов с пропущенными значениями в рассматриваемой задаче не влечет сильной потери данных, кроме того, такой метод часто применяется в предобработке данных в машинном обучении, по этой причине было решено использовать именно его для борьбы с пропущенными значениями.

Итоговый набор данных был представлен в виде матрицы признак × объект размером 7 × 2803, где объект – слог. Вектор признаков состоял из 6 параметров:

- log f min – минимальная частота основного тона;
- log f max – максимальная частота основного тона;
- log f mean – средняя частота основного тона;
- log int min – минимальная интенсивность;
- log int max – максимальная интенсивность;
- log int mean – средняя интенсивность.

Верификация личности

В была рассмотрена верификация личности как задача бинарной классификации. В датасете целевая переменная была представлена 0 или 1, где 0 – совокупность объектов и признаков, принадлежащих всем дикторам базы, кроме одного верифицируемого, объектам класса которого принадлежала метка 1.

Таб. 1

	F1	*Precision*	*Recall*	*Accuracy*
Baseline	0.04	0.02	1	0.02
Balanced RF	0.157	0.087	0.821	0.707
Balanced bagging	0.177	0.102	0.678	0.790
RF SMOTE	**0.258**	**0.184**	0.428	0.917
RF ADASYN	0.252	0.179	0.428	0.915

Поставленная в данной работе задача бинарной классификации имеет значительное ограничение – несбалансированность классов. Представление одного класса как совокупности признаков одного говорящего, а второго как совокупности признаков остальных 49 дикторов означает превосходство количества объектов одного класса над другим более чем в 10 раз.

Для решения проблемы несбалансированности классов в задаче верификации личности были использованы oversampling-методы SMOTE, ADASYN к классификатору Random Forest, а также самостоятельные классификаторы Balanced RF и Balanced bagging. Результаты классификации представлены в таблице 1.

По итогам применения данных методик к задаче верификации лучшее качество на трех метриках, включая F1, достигается с помощью метода Random Forest + SMOTE (F1 = 0.258).

Кластеризация данных

Для лучшего понимания структуры данных был применен метод статистики Хопкинса, оценивающий кластерные свойства датасета, а также метод кластеризации k-средних (k-Means). Число кластеров определялось методом силуэта. Так, наиболее оптимальное число кластеров по силуэтному анализу – 2 (2 clusters: 0.38), большему числу кластеров соответствовал меньший коэффициент. При этом значение статистики Хопкинса составило 0.9, что говорит о существовании четкой кластерной структуры данного датасета. Таким образом, в анализируемом наборе прослеживается четкая кластерная структура, однако в основе ее деления не лежит многоклассовая сущность набора данных, а какой-либо другой принцип, выделяющий 2 кластера, а не 50. Предположительно, такое деление может быть связано с гендерным разделением говорящих. Проверка данной гипотезы описана в пункте ниже.

Установление пола говорящего

Проведенный ранее анализ кластерной структуры данных показал, что данный набор признаков действительно имеет четкую кластерную структуру, а значит, показатель в 2-3 кластера не случаен. На этом основании целесообразно предположить, что такое выделение кластеров связано с гендерной принадлежностью дикторов: 50% выборки – лица мужского пола, 50% - женского. Кроме того, из предшествующих исследований (Р.К. Потапова, L. Mary) известно, что просодика, а именно мелодические признаки сохраняют информацию, уникальную для представителей разных гендеров.

Таким образом, в рамках данной части исследования было решено применить методы машинного обучения для задачи идентификации пола говорящего по просодическим параметрам как подзадачи идентификации/верификации. Как и в предыдущем пункте, задача сводилась к бинарной классификации. Результаты качества на тестовой выборке по методам классификации представлены в таблице 2.

Таб. 2

	F1	*Precision*	*Recall*	*Accuracy*
DTrees	0.96	0.96	0.96	0.96
RForest	**0.97**	**0.97**	**0.97**	**0.97**
kNN	0.96	0.96	0.96	0.96

Таким образом, следует заключить, что модель Random Forest с числом деревьев, равным 50, достигает максимального качества в задаче идентификации пола говорящего по просодическим параметрам, позволяя с точностью в 98% определить его пол.

Как уже отмечалось ранее, кластерная структура данных четко прослеживается, разграничивая 2 основных подмножества. Это позволяет предположить, снижение размерности до двух- и трехмерного пространства векторов даст возможность наглядно проиллюстрировать 2 выделенных кластера в соответствующих пространствах (рис. 1 и 2). Техника понижения размерности PCA позволяет визуализировать данные.

Рис. 1. Визуализация кластерной структуры данных в двухмерном пространстве.

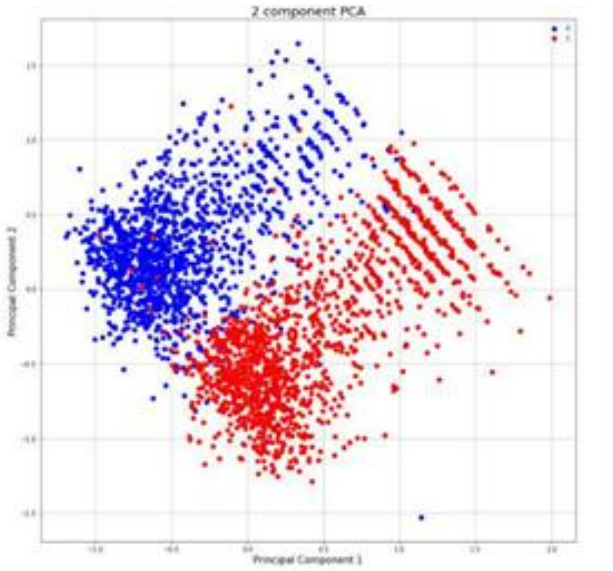
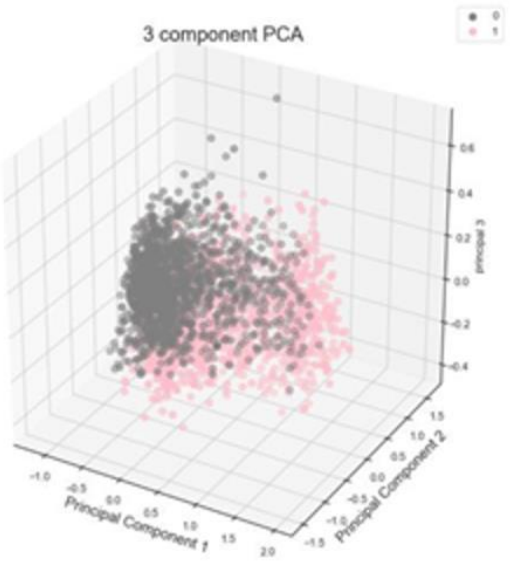


Рис. 2. Визуализация кластерной структуры данных в трехмерном пространстве.



Основные выводы

Показатель точности верификации F1 = 0.258 с применением техники Random Forest+SMOTE, несмотря на то, что не может лежать в основе реальных верификационных систем, тем не менее, подчеркивает значимость просодики в распознавании личности по голосу. Исходя из данного показателя, мы можем предположить, что хотя просодика в одиночку и не способна быть решающей методикой в идентификации личности, потенциальное увеличение качества кепстральной системы за счет просодики может оказаться значительным.

Кроме того, точность установления пола говорящего в 98% говорит о том, что просодика способна эффективно решать задачу гендерной принадлежности диктора как подзадачу верификации диктора.

Дальнейшее развитие данного исследования может иметь несколько направлений. Во-первых, может быть рассмотрено объединение просодических признаков с мел-кепстральными коэффициентами с целью создания единой качественной верификационной системы. Во-вторых, представление слогов как объектов просодической системы в виде единой последовательности станет возможным при применении искусственных нейронных сетей как метода классификации. Также, вместо логарифмирования просодических признаков, в качестве их нормализации могут быть использованы Гауссовские смеси.

Библиография

1. Adami A. (2005), Prosodic modeling for speaker recognition based on sub-band energy temporal trajectories, In Proceedings of the ICASSP, Vol. 1, pp. 189–192.
2. Boersma P., Weenink D. (1992), Praat: Doing phonetics by computer available at: www.praat.org
3. Campbell J.P. (1997), Speaker recognition: A tutorial, In Proceedings of the IEEE, Vol. 85, pp. 1437– 1462.

4. Chawla N., Bowyer K., Hall L., Kegelmeyer W. (2002), SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Vol. 16, pp. 321–357
5. Farru's M. (2008), Fusing Prosodic and Acoustic Information for Speaker recognition, PhD

   Dissertation, TALP Research Center, Speech Processing Group Department of Signal Theory and Communications Universitat Politecnica de Catalunya, Barcelona.
6. John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. (1993), TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium.
7. Markel J.D., Davis S.B. (1978)б Text-independent speaker identification from a large linguistically unconstrained time-spaced data base, In Proceedings of the ICASSP, Vol. 3, pp. 287– 290.
8. Mary L., Yegnanarayana B. (2006), Prosodic features for speaker verification, In Proceedings of Interspeech, Vol. 2, pp. 917–920.
9. Mary L., Yegnanarayana B. (2008), Extraction and representation of prosodic features for language and speaker recognition, Speech Communication, Vol. 50, pp. 782–796.
10. Mary L. (2019), Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition, Springer, London, UK.
11. Muller A. (1997), The power of voice [Die Macht der Stimme], Tubingen, Germany.
12. Potapova R. K. (2002), New informational technologies and linguistics [Novyie informacionnyie technologii i lingvistika], M, Russia, Moscow.
13. Schiel F. (1999), Automatic Phonetic Transcription of Non-Prompted Speech, Proc. of the ICPhS, San Francisco, pp. 607-610.
14. Shriberg E., Ferrer L., Kajarekar S., Venkataraman A. Stolcke A. (2005), Modeling prosodic feature sequences for speaker recognition, Speech Communication, Vol. 46, pp. 455–472.
15. Sönmez K., Shriberg E., Heck L., Weintraub M. (1998), Modeling dynamic prosodic variation for speaker verification. In: Proc. Internat. Conf. on Spoken Language Processing, pp. 3189–3192.
16. Sorokin V.N., Makarov I.S. (2008), Gender recognition from vocal source, Acoustical Physics, Vol. 54, pp. 571-578.
17. Wolf J. J. (1972), Efficient acoustic parameters for speaker recognition, Journal of the Acoustical Society of America, Vol. 51, pp. 2044–2056.