Неявная модель произношения для автоматического распознавания речи

Владимир Яковлевич Чучупал,
в.н.с., Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

Аннотация

Вариативность произнесения слов в естественной разговорной речи является одним из основных источников ошибок при ее автоматическом распознавании. Примером подобной вариативности является пропуск или подмена отдельных звуков, вызванная неполной или нечеткой артикуляцией в быстрой речи.
В статье описана неявная модель произношения, которая реализована посредством сглаживания параметров акустических моделей соседних звуков.
Предлагается использовать контексто-зависимые параметры сглаживания, которые обусловлены текущим фонетическим, просодическим и языковым контекстом звуков.
Хотя подход к моделированию вариативности произношения уже обсуждался в литературе, метод контексто-зависимого сглаживания моделей смежных звуков, насколько известно автору, пока не был представлен.
Эксперименты на речевом корпусе данных, который содержал как читаемую, так и естественную речь, показали корректность предложения использовать переменные параметры сглаживания, значение которых обусловлено фонетическим и просодическим контекстом.

Ключевые слова: автоматическое распознавание речи, обработка естественного языка, акустическое моделирование речи, модели вариативности произношения

ВВЕДЕНИЕ

Это было подтверждено симуляционными экспериментами ~\cite {Saraclar Nock,McAllaster} в которых использование корректных фактических транскрипций вместо канонических привело к снижению уровня ошибок распознавания почти в два раза.

Существует два основных подхода к моделированию вариативности произношения. Явный подход описывает вариабельность произнесения слова путем описания возможных изменений в базовой фонемной транскрипции ~\cite {Wester} слов.

Неявный подход ~\cite {Saraclar} описывает изменчивость произношения посредством изменений в структуре акустических моделей звуков, не изменяя фонемные транскрипции.

Системы распознавания речи обычно реализуют явный подход, поскольку он выглядят естественнее и может быть просто описан в терминах классической модели распознавания речи.

Пусть $X = \{x^t\}$, $t = 1, ..., T$ последовательность наблюдений, векторов речевых параметров, а $W = \{w_i\}$, $i = 1, ..., N$ последовательность слов. Наиболее вероятная последовательность слов W^* при известных X может быть получена из следующего выражения [5]:

$$\begin{aligned} W^* &= \arg\max_W P(W|X) \\ &= \arg\max_W \frac{P(X|W)P(W)}{P(X)} \\ &= \arg\max_W P(X|W)P(W). \end{aligned} \quad (1).$$

Первый сомножитель $P(X|W)$ в числителе (1) соответствует правдоподобию заданных наблюдений и вычисляется с помощью акустической модели. Вероятность последовательности слов $P(W)$ вычисляется с помощью модели языка. Знаменатель $P(X)$, по сути, является нормализующим членом.

Обозначим транскрипцию слова w как t^w, множество из всех транскрипций слова w обозначим как T^w. Множество всех возможных транскрипций для последовательности слов W обозначим через T^w. Обозначение t^w будет использовано для обозначения любого элемента T^W. Тогда условие (1) можно аппроксимировать (аппроксимация Витерби) следующим выражением:

$$W^* = \arg\max_{W, t^W \in T^W} P(X|t^W)P(t^W|W)P(W). \quad (2).$$

Оценка величины $(t^W|W)$ выполняется на основе модели вариативности произношения. Параметрами этой модели являются фонемные транскрипции из T^W, а также их относительные частоты $\{P(t^W|W), t^W \mid T^W\}$. Таким образом, явная модель вариативности произношения может быть описана на основе методов определения фонемных транскрипций слов и их вероятностей.

Наиболее очевидный способ выбора возможных фонемных транскрипций основан на использовании фонемного транскриптора, т.е. системы распознавания фонем в потоке речи. На практике это пока недостаточно эффективно из-за низкой точности существующих транскрипторов. В качестве оценки вероятности фонемных транскрипций естественно использовать их экспериментальные относительные частоты. Однако это часто не представляется возможным, поскольку требует очень больших корпусов речевых данных.

В литературе описаны способы преодоления описанных трудностей, которые, в частности, основаны на дискриминантном и непараметрическом подходах [3, 6, 7, 8, 9], однако выигрыш в пословной точности распознавания речи, который получается при использовании явных моделей произношения, далеко не так существенен, как этого можно было бы ожидать, исходя из результатов симуляционных экспериментов.

Основная идея явного моделирования заключается в том, что все возможные изменения произношения могут быть достаточно точно представлены изменениями в базовой фонемной транскрипции слов, т.е. заменами, вставками и удалениями фонем. Экспериментальный анализ показывает, что альтернативное описание изменчивости произношения, особенно в спонтанной речи, может быть сделано на основе использования моделей, которые способны представлять частичные изменения фонемного качества, т.е. путем неявного моделирования произношения [4].

Модели неявного моделирования вариативности произношения в отличие от явных реализуются на основе более сложных акустических моделей фонем, которые используются в базовых фонемных транскрипциях. Например, марковских моделей в виде сети из состояний. Тем не менее, как следует из [4, 11, 12, 13, 14, 15] результат использования неявных моделей — повышение точности распознавания — существенно не отличается от такового для явных моделей.

Опубликованы исследования, в которых показано, что уровень вариативности произношения часто имеет условный характер [16], т.е. зависит от окружающего контекста. Это обстоятельство можно интерпретировать как возможность предсказать уровень вариативности произношения слов на основе текущих просодических, синтаксических и семантических характеристик речи.

КОНТЕКСТО-ЗАВИСИМАЯ НЕЯВНАЯ МОДЕЛЬ ВАРИАТИВНОСТИ ПРОИЗНОШЕНИЯ

Пусть m и $\$n\$$ обозначают модели звуков, $P(x|m), P(x|n)$ обозначают условные вероятности для наблюдений речевых параметров $\$x\$$ при заданных моделях. Тогда модель m, которая сглаживается моделью n, $P(x|m,n)$, может быть определена аналогично [14]:

$$P_\lambda(x|m,n) = \lambda * P(x|m) + (1-\lambda)P(x|n), \quad 0 \le \lambda \le 1. \tag{3}$$

Выражение (2) позволяет описать, пусть в упрощенной форме, некоторые часто наблюдаемые в спонтанной речи произносительные изменения.

Например, значение λ, равное 0, означает, что звук m полностью заменен звуком $\$n\$$. Аналогично λ = 0,5 описывает ситуацию неполной замены фонетического качества, которая упрощенно соответствует эффектам назализации, озвончения или оглушения звуков в спонтанной речи.

Пусть звук m наблюдается на временном интервале $s(m)$, ..., $e(m)$ с соответствующими параметрами $x_{s(m)}, ..., x_{e(m)}$. Оптимальное значение коэффициента сглаживания λ в (2) может быть найдено аналогично оценке оптимальных весов смесей при обучении модели смеси нормальных распределений [17]:

$$\lambda_{m,n} = \frac{\sum_{t=s(m)}^{e(m)} P(x_t|m)}{\sum_{t=s(m)}^{e(m)} (P(x_t|m) + P(x_t|n))}. \tag{4}$$

Для корпуса данных U, который состоит из R высказываний $U = \{u^r | r = 1, ..., R\}$, значение $\lambda_{m,n}$ может быть вычислено усреднением локальных оценок (4) для всех наблюдений пар фонем (m,n) аналогично оценкам параметров марковских моделей [17]?

$$\hat{\lambda}_{(m,n)} = \frac{\sum_{r=1}^{R} \sum_{(m,n)\in u_r} \lambda_{(m,n)} P(m)}{\sum_{r=1}^{R} \sum_{(m,n)\in u_r} P(m)}, \quad (5)$$

где $P(m)$ правдоподобие звука m при наблюдении (m,n).

Пусть V — это вектор контекстных признаков, то есть признаков фонемного, позиционного, просодического, лингвистического и синтаксического контекста, наличие которых [16] коррелирует с наблюдаемым уровнем произносительной вариативности:

$$V\,(c,\ l,\ r;\ nPh,\ pPOS,\ ROS,\ wPOS,\ POS,\ eWrd,\ LM):$$

where

c: центральная фонема,
l: левый фонемный контекст,
r: правый фонемный контекст,
nPh: следующая фонема,
$pPOS$: позиция в слове,
ROS: темп речи,
$wPOS$: позиция слова во фразе,
POS: часть речи слова,
LM: значение модели языка. (6)

Экспериментально проверим предположение о том, что уровень произносительной вариативности, описываемый значением $P(\lambda|V)$, действительно зависит от признаков (6).

ЧИСЛЕННОЕ ИССЛЕДОВАНИЕ

Доказательство того, что модели звуков (2) с использованием интерполированных параметров могут существенно снизить уровень ошибок, лучше всего сделать на основе эксперимента по распознаванию речи. Однако такой эксперимент требует внедрения интерполированных моделей в процедуры обучения и распознавания.

Предварительные эксперименты, которые доказывают эффективность предложенных моделей, пусть и косвенно, могут быть выполнены путем оценки параметра λ на корпусе данных, чтобы показать, что значение параметра существенно зависит от контекстных признаков (6).

Для экспериментальных исследований использовался речевой материал из корпусов данных с русской устной речью: TeCoRus [18], RuSpeech [19] и PronExRu [20].

Речевой материал был разделен на три выборки, предназначенные для обучения, настройки и тестирования. Обучающая выборка состояла из речевого материала корпусов RuSpeech и TeCoRus, которые содержали читаемую речь от 200 дикторов. Настроечные данные использовались для оценки значений параметров интерполяции. Они состояли из тестового материала корпуса RuSpeech, всего 1000 высказываний от 10 человек. Наконец, тестовый материал состоял из данных PronExRu (200 высказываний, 9 человек).

Обучающие данные использовались для создания гендерно-ориентированных акустических моделей. Оценка значений признаков в (6), за исключением признака темпа речи ROS (rate of speech), выполнялась двумя этапами с использованием результатов автоматического распознавания речи, полученных на первом этапе.

Оценка параметра ROS была выполнена с использованием алгоритма [13]. Чтобы понизить вычислительную сложность, вместо полного перебора возможных комбинаций длительностей звуков использовалась процедура сэмплинга длительностей.

В ходе предварительных экспериментов дискретные значения признаков были оценены для каждого центрального состояния марковских моделей. Затем значения параметров λ были вычислены с использованием (5).

Чтобы иметь возможность ранжировать признаки, строилось два бинарных дерева решений. Одно дерево было построено с использованием читаемого речевого материала корпуса TeCoRus. Другое дерево было построено с использованием спонтанной речи корпуса данных PronEx.

Набор вопросов для выращивания деревьев содержал всевозможные вопросы о наличии признаков (6) для текущего состояния (например: «Применяется ли это состояние к существительному?», «Является ли это состояние частью слова?»), кроме того, вопросник включал набор стандартных вопросов, которые используются при построении фонемных бинарных деревьев решений для синтеза алфавита контекстo-зависимых состояний марковских моделей аллофонов. Всего использовалось 82 вопроса: 30 вопросов, относящихся к признакам из (6), и 52 вопроса, относящиеся к фонемному контексту звука и его фонемному качеству.

В качестве критерия расщепления вершин при выращивании деревьев использовалась величина информационного выигрыша [21], изменения энтропии при разделении родительской вершины на две дочерние после использования вопроса:

$$\Delta H_\lambda(t,f) = H_\lambda(t) - \frac{|t_+|}{|t|}H_\lambda(t_+) - \frac{|t_-|}{|t|}H_\lambda(t_+). \qquad (7)$$

Здесь $H_\lambda(t)$, $H_\lambda(t\_)$, $H_\lambda(t_+)$ обозначают величину энтропии для данных в родительской вершине t и энтропию данных в дочерних вершинах $t\_$ и t_+, которые образовались после бинарного разделения родительской вершины t.

Чтобы численно оценить важность каждого признака, рассчитывались значения весов признаков. Для этого для каждого данного признака по всем вершинам дерева, где вопрос об этом признаке был выбран как лучший разделитель вершины, значения величины информационного выигрыша суммировались. То есть вычислялось суммарное изменение выигрыша (7) для данного признака. Затем эти суммарные изменения нормировались величиной максимального значения к 100 (т.е. самый важный признак имел вес 100):

$$I(f) = \sum_{t \in T} \Delta H_\lambda(t,f), \qquad (8)$$

где суммирование выполняется по всем вершинам дерева.

Результат построения деревьев содержится в таблице 1, которая содержит список наиболее важных признаков для настроечной и тестовой частей корпуса данных.

Таблица 1

Наиболее важные признаки вариативности

Признак	Вес для настроечных данных	Признак	Вес для тестовых данных
ROS-noun	100	RATE-F	100
RATE-MDL	72	L-Labial	70
L-Soft	55	STRESS	63
RATE-F	42	POS-noun	54
R-VoiceLess	40	L-VOW	52
L-Sonant	37	PpMdl	51
R-Forv	35	R-Labial	42
L-Forv	31	NEXT-VOW	39
POS-adj	30	RATE-FST	39
R-Sil	29	R-Forv	35
STRESS	29	R-Forw	29

Сокращения в названиях признаков имеют следующее значение:

"POS-noun" обозначает «существительное»,
"POS-adj" обозначает «имя прилагательное»,
"RATE-F" обозначает «быстрый темп речи»,
"RATE-mdl" обозначает «умеренный темп речи»,
"STRESS" обозначает «ударный звук»,
"PpMdl" обозначает «звук в середине слова»,
"R-sil обозначает «звук после паузы»,
"L-soft" обозначает «предыдущий звук — мягкий согласный»,
"R-VoiceLess" обозначает «звук перед паузой»,
"L-Sonant" обозначает «предыдущий звук звонкий согласный»,
"R-Forv" обозначает «следующий звук согласный передний»,
"R-Forw" обозначает «следующий звук — передний гласный»,
"NEXT-VOW" обозначает «следующий звук — гласный»,
"R-Labial" обозначает «следующий звук — губной согласный».

Значения сглаживающего параметра λ, измеренные с помощью (4), большую часть времени находились в диапазоне от 0,2 до 0,8, и значение правдоподобия данных для сглаженных моделей было выше, чем при исходных акустических моделях.

Как следует из таблицы 1, наиболее важным признаком для читаемого речевого материала (речевого корпуса RuSpeech) является признак части речи. Для спонтанной речи (корпус Pronex) наиболее важным признаком является признак темпа речи. Оба эти признака относятся к предлагаемому набору признаков вариативности произношения (6). Из одиннадцати наиболее важных признаков для материала читаемой речи, как следует из таблицы, 6 признаков, которые входят в набор (6). Для материала спонтанной речи таких признаков 5.

Если рассмотреть наиболее важные признаки с весом более 50, то для читаемой речи таких признаков всего три, и два из них относятся к набору (6). Для спонтанной речи таких признаков шесть, причем четыре из них из набора (6).

Наиболее важными для обоих корпусов данных являются признаки, связанные с темпом речи, частью речи и положением заука относительно ударения. Исходя из этих предварительных результатов, можно утверждать, что значение λ, определенное в соответствии с (5), действительно зависит от значений признаков, перечисленных в (6). Предлагаемые признаки вариативности произношения, основанные на просодических и позиционных контекстах слова и звука, коррелируют с наблюдаемыми изменениями акустического качества, т.е. с изменчивостью произношения.

Таким образом, модель (3), основанная на сглаживании параметров соседних акустических моделей, может использоваться для учета эффектов изменчивости произношения при распознавании естественной речи. Такая модель может быть реализована, например, путем организации пост-обработки как второй проход при распознавании.

ЗАКЛЮЧЕНИЕ

В статье предложена неявная модель вариативности произношения как способ повышения точности автоматического распознавания речи. Произносительные изменения в разговорной речи предлагается учитывать за счет использования акустических моделей звуков со сглаженными параметрами, так, что величина параметров, которые регулируют уровень, зависит от контекстных и позиционных признаков звука и слова, в котором он находится.

Предложен набор из нескольких потенциальных контексто-зависимых признаков вариативности и численно исследована их ценность как предсказателей вариативности. Для этого на материале читаемой и спонтанной речи с использованием вопросника, который содержал как стандартные для построения фонетических решающих деревьев вопросы, так и вопросы, относящиеся к потенциальным признакам вариативности, были построены решающие деревья. Все признаки были ранжированы по важности в соответствии с метрикой информационного выигрыша.

Было установлено, что ряд предложенных контексто-зависимых признаков вариативности фактически коррелирует с наблюдаемой изменчивостью произношения, превосходя по важности большинство обычных признаков фонемного контекста звука. Таким образом, предложенная модель, основанная на сглаживании параметров, может быть использована для учета эффектов изменчивости произношения в естественном распознавании речи.

Список литературы

1. *Saraclar M., Nock H., Khudanpur S.* Pronunciation modeling by sharing Gaussian densities across phonetic models // Computer Speech and Language. 2000. Vol. 14(4). P. 137–160.
2. *McAllaster D., Gillick L., Scattone F., Newman M.* Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. // Int.Conf. Speech and Language Processing, 1998, Sydney, P. 1847–1850.

3. *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // Computer Speech and Language, 2003. Vol. 17, P. 69–85.
4. *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // Computer Speech and Language 2004. Vol. 18(4). P. 375–395.
5. *Jelinek F.* Statistical Methods for Speech Recognition // The MIT Press, Cambridge, Massachusetts, 1997.
6. *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition. // Proc.Conf.of International Speech Communication Association, Interspeech, 2014, Singapoure, pp. 1458-1462, 2014.
7. *Byrne B., Finke M., Khudanpur S., McDonough J., Nock H., Riley M., Saraclar M., Wooters C., Zavaliagkos G.* Pronunciation modelling for conversational speech recognition: a status report from WS97. // IEEE Workshop on Automatic Speech Recognition and Understanding. 1997, USA, P. 26–33. 10.1109 ASRU.1997.659004
8. *Hitchinson B., Droppo J.* Learning Non-Parametric Models of Pronunciation in automatic speech recognition // Proc. International Conference on Acoustics, Speech, and Signal Processing, ICASSP, USA, 2011, P. 4904–4907.
9. *Schramm H.* Modeling Spontaneous Speech Variability for Large Vocabulary Continuous Speech Recognition. // Doktors der Naturwissenschaften Dissertation, Technical University of Aachen, Germany, 2006.
10. *Livescu L., Glass J.* Feature-based pronunciation modeling for speech recognition // Proc. Human Lang. Tech. Conf. of the North American Chapter of the Assoc. for Comp. Ling., USA, 2004.
11. *Hain T., Woodland P.C.* Dynamic HMM selection for continuous speech recognition // Proc. of EuroSpeech, 1999. P. 1327–1330.
12. *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // Speech Communication. Vol 46 (2005). P. 171–188.
13. *Zheng J., Franco H., Stolcke A.* Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition // Speech Communication. 2003. Vol. 41. P. 273–285.
14. *Liu Y.* Modeling partial pronunciation variations for spontaneous Mandarin speech recognition // Computer Speech \& Language. Vol. 17. No 4, 2003. P. 357–379.
15. *Spiess T., Wrede B., Fink G.A., Kummert F.* Data-driven Pronunciation Modeling for ASR using Acoustic Subword Units // Int Conf. InterSpeech, 2003. P. 2549–2552.
16. *Ostendorf M., Shafran I., Bates R.* Prosody Models For Conversational Speech Recognition // 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, USA, 2003. P. 147–154.
17. *Rabiner L., Biing-Hwang J.* Fundamentals of Speech Recognition.// PTR, Prentice Hall Signal Processing Series, New Jersey, USA, 1993.
18. *Чучупал В.Я., Маковкин К.А, Чичагов А.В., Кузнецов В.Б., Огарышев В.Ф.* Речевой корпус данных TeCoRus // Свидетельство об официальной регистрации базы данных №2005620205, 2005 г.
19. *Arlazarov, V.L., Bogdanov, D.S., Krivnova, O.F., Podrabinovitch, A.Ya.* Creation of Russian Speech Databases: Design, Processing, Development Tools, . Proceedings of the Intern. Conference SPECOM'2004, 4Pp. 650–656 , S-Pb., Russia, 2004.
20. Russian spoken speech corpus //Database registration certificate~2016620687, Rospatent, Moscow, 2016 (in Russian).
21. *Quinlan, J.R.* Induction of decision trees. // Machine Learning 1, 81–106, 1986.

IMPLICIT PRONUNCIATION VARIATION MODEL FOR AUTOMATIC SPEECH RECOGNITION

Vladimir J. Chuchupal,
leading scientific researcher, Computing center. A. A. Dorodnicyn FITS Yiwu wounds

Abstract

The variations in pronunciation of words in natural speech are one of the main sources of automatic speech recognition errors. The examples of such variations include the pronunciation variations that are caused by a fuzzy or an incomplete articulation that is frequently observed in spontaneous speech.

The implicit pronunciation model is proposed that is implemented by means of smoothing of parameters of the adjacent acoustical phone models in phonemic transcription. It is proposed to use the context-dependent smoothing, so that the values of the smoothing parameters are conditioned by the current position and prosodic contexts of a phone.

While the pronunciation variation modeling approach on the base of combination of acoustical models has already been discussed in literature, the method based on the context-dependend smoothing of the adjacent models as far as we know has not been published yet.

The experiments on the speech corpuses that contained both the read and spontaneous speech showed the correctness of the proposal for the use of the context-dependent smoothing parameters which are conditioned by the features of phonemic context and prosody.

Keywords: automatic speech recognition, natural speech processing, acoustic modeling, pronunciation variation modeling