Неявная модель вариативности произношения для автоматического распознавания речи*

В. Я. Чучупал
v.chuchupal@gmail.com
ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Вариативность произношения слов и словосочетаний в естественной разговорной речи является одним из основных источников ошибок при автоматическом распознавании речи, поэтому использование моделей вариативности произношения представляется важным направлением повышения эффективности работы систем распознавания речи. Рассматривается проблема моделирования вариативности, которая вызвана нечеткой, неполной артикуляцией, например в результате нарушения синхронизации работы органов речеобразования. Предлагается использовать неявные произносительные модели, основанные на комбинировании акустических моделей соседних звуков. Комбинирование может заключаться в сглаживании или интерполяции параметров акустических моделей текущих звуков параметрами соседних моделей. Степень проявления вариативности, вообще говоря, зависит от синтаксического и просодического контекста звука, поэтому предлагается использовать меняющиеся параметры интерполяции в зависимости от наличия позиционных, фонетических, синтаксических и просодических признаков. Подход с комбинированием акустических моделей на основе сглаживания их параметров был описан в научной литературе, однако автору неизвестны исследования с комбинированием именно соседних акустических моделей, где бы параметры комбинирования зависели от текущих контекстных признаков. Предварительные эксперименты на корпусах с читаемой и разговорной речью показали справедливость предположений о целесообразности использования интерполяционных моделей и существования зависимости параметров сглаживания моделей от наличия позиционных и синтаксических признаков.

Ключевые слова: *автоматическое распознавание речи; вариативность речи; моделирование произношения; скрытые марковские модели*

DOI: 10.21469/22233792.2.4.01

1 Введение

Использование моделей вариативности произношения имеет высокий потенциал как способ повышения эффективности автоматического распознавания речи. Это подтверждается данными так называемых симуляционных экспериментов, когда за счет использования корректных фактических произносительных транскрипций уровень пословной ошибки распознавания — WER (word error rate [1]) — понижался в разы [2, 3].

В литературе встречаются два основных подхода к моделированию вариативности произношения. Явное моделирование (explicit modeling) заключается в моделировании вариативности произнесения путем описания возможных изменений в фонемной транскрипции слов [4]. Неявное моделирование (implicit modeling) [5] описывает вариативность произнесения путем изменений в структуре моделей звуков в канонической транскрипции слов. В прикладных системах обычно используется явное моделирование, которое естественным

*Работа выполнена при финансовой поддержке РФФИ, проект № 14-01-00607.

образом описывается в рамках классической статистической формулировки распознавания слитной речи. Если $X = \{x_t\}, t = 1, \ldots, T$, — наблюдаемый образ в виде последовательности параметров речевого сигнала, а $W = \{w_i\}, i = 1, \ldots, N$, — последовательность слов словаря, то результат распознавания X в виде наиболее вероятной последовательности слов W^* определятся из уравнения [6]:

$$W^* = \arg\max_W P(W|X) = \arg\max_W \frac{P(X|W)P(W)}{P(X)} = \arg\max_W P(X|W)P(W). \qquad (1)$$

Первый сомножитель $P(X|W)$ в числителе (1) соответствует правдоподобию данных при заданной последовательности слов и определяется с помощью акустических моделей. Полученная величина правдоподобия затем умножается на значение $P(W)$, которое определяется с помощью модели языка. Знаменатель $P(X)$ — вероятность наблюдения X, выполняет функции нормализующего члена.

Обозначим фонемную транскрипцию слова w через t^w, множество всех фонемных транскрипций этого слова обозначим T^w. Множество возможных транскрипций последовательности слов W обозначим T^W. Запись t^W будет использоваться для обозначения какой-либо конкретной последовательности транскрипций из T^W. Тогда уравнение (1) можно аппроксимировать (так называемая аппроксимация Витерби) выражением:

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

Оценка $P(t^W|W)$ осуществляется моделью вариативности произношения, параметрами которой являются фонемные транскрипции T^W и условные вероятности их реализации $\{P(t^W|W), t^W \in T^W\}$.

Таким образом, явные модели вариативности можно идентифицировать на основе используемых методов выбора фактических фонемных транскрипций и определения их вероятностей. Сложность заключается в том, что наиболее очевидный способ выбора фактических транскрипций с помощью фонемного распознавателя до недавнего времени был неэффективен ввиду низкой точности таких распознавателей, а использование естественных частотных оценок для вероятностей реализации транскрипций затруднительно по причине отсутствия корпусов данных требуемого размера.

Различные способы преодоления этих проблем достаточно широко описаны в литературе, но полученный за счет использования явных моделей выигрыш в величине WER для разговорной речи не так велик, как можно было ожидать: 0,8% [4, 7], 2,2% [8], 1,8% [9], 0,9% [10].

Основой явных моделей вариативности является предположение, что произносительные изменения в разговорной речи можно достаточно адекватно описать полными заменами (включая вставку и удаление) фонем. В то же время анализ экспериментальных данных показывает [5], что более точным описанием вариативности произношения, особенно в типичной ситуации, когда вариативность вызвана нарушением синхронизации движений речевых органов [11], является использование моделей, способных представлять неполные изменения фонемного качества звуков. Такие произносительные модели относят к так называемым неявным.

В отличие от явных моделей неявные модели реализуются как часть акустических моделей, например за счет усложнения их структуры либо использования множественных моделей. Практически выигрыш в точности распознавания для неявных моделей не отличается существенно от такового же для явных в терминах величины пословной ошибки:

1,7% [5], 0,7% [12, 13], 2,2% [14], 2,39% (послоговой ошибки для китайского языка) [15], 2,5% [16].

Несмотря на наличие исследований в пользу условного характера проявления вариативности в разговорной речи [17], которые можно интерпретировать как возможность предсказать появление вариативности исходя из синтаксических и семантических характеристик речевого сигнала, в литературе мало конкретных результатов в этом направлении.

2 Моделирование вариативности произношения путем сглаживания параметров акустических моделей

Пусть m и n обозначают звуки, а $P(x|m), P(x|n)$ — их акустические модели, которые определяют условные вероятности наблюдения параметров x. Интерполированную (или сглаженную) модель для m, $P(x|m,n)$ определим как [15]:

$$P_\lambda(x|m,n) = \lambda P(x|m) + (1-\lambda)P(x|n), 0 \leqslant \lambda \leqslant 1. \tag{2}$$

Форма (2) позволяет упрощенно описать некоторые частые эффекты вариативности произношения. Например, значение коэффициента $\lambda = 0$ означает, что звук m пропущен, а $\lambda = 0{,}5$ соответствует частичному изменению его качества, например оглушению, озвончению, назализации, если n обладает этими признаками.

Если реализация звука m соответствует параметрам $x_{s(m)}, \ldots, x_{e(m)}$ на отрезке времени $s(m), \ldots, e(m)$, то наиболее правдоподобная оценка коэффициента λ для (2) вычисляется аналогично оценке весов смесей при обучении GMM (gaussian mixture models) моделей [18]:

$$\lambda_{m,n} = \frac{\sum_{t=s(m)}^{e(m)} P(x_t|m)}{\sum_{t=s(m)}^{e(m)} (P(x_t|m) + P(x_t|n))}. \tag{3}$$

Для корпуса данных из R реализаций (предложений) $U = \{u_r | r = 1, \ldots, R\}$ с, вообще говоря, несколькими произносительными вариантами, транскрипциями $\{f_r | r = 1, \ldots, F_r\}$, значение параметра $\hat{\lambda}_{m,n}$ можно вычислить, усредняя локальные значения (3) по всем вхождениям пар звуков (m,n):

$$\hat{\lambda}_{(m,n)} = \frac{\sum_{r=1}^{R} \sum_{f=1}^{F_r} \sum_{(m,n)} \lambda_{(m,n)} P(m)}{\sum_{r=1}^{R} \sum_{f=1}^{F_r} \sum_{(m,n)} P(m)}, \tag{4}$$

где $P(m)$ — правдоподобие звука m для вхождения (m,n).

Поскольку модель (2) не использует никакой информации, кроме параметров соседних (в экспериментах n была правым контекстом m) моделей, учитывая результаты [15], можно ожидать, что заметного выигрыша от ее использования не будет.

Определим вектор признаков вариативности как вектор V, составленный из контекстных признаков, которые вероятно коррелируют с проявлениями вариативности произношения [17]: $V(c, l, r, \mathrm{nPh}, \mathrm{pPOS}, \mathrm{ROS}, \mathrm{wPOS}, \mathrm{POS}, \mathrm{mWrd}, \mathrm{fWrd}, \mathrm{LM})$, где

$$\left.\begin{array}{rl} c: & \text{центр;} \\ l: & \text{левый контекст;} \\ r: & \text{правый контекст;} \\ \text{nPh}: & \text{следующая фонема;} \\ \text{pPOS}: & \text{позиция фонемы;} \\ \text{ROS}: & \text{темп речи;} \\ \text{wPOS}: & \text{позиция слова;} \\ \text{POS}: & \text{часть речи слова;} \\ \text{mWrd}: & \text{словосочетание;} \\ \text{fWrd}: & \text{частотность слова;} \\ \text{LM}: & \text{значение модели языка.} \end{array}\right\} \quad (5)$$

Проверим предположение о зависимости степени вариативности произношения от наличия признаков из набора (5) и, в случае положительного ответа, построения интерполяционной модели в форме (2), которая может быть использована при распознавании речи.

3 Экспериментальное подтверждение эффективности модели вариативности произношения

Проверка утверждения, что использование интерполированных моделей вариативности произношения эффективно с точки зрения повышения точности автоматического распознавания речи очевидно должна проводиться в рамках экспериментов по распознаванию. Проведение таких экспериментов требует встраивания интерполированных моделей в существующий программный код для оценки параметров и распознавания, существенной модификации программного обеспечения. На данном этапе работы проверка эффективности осуществляется косвенными методами, т. е. путем вычисления и анализа значений параметра интерполяции λ на тестовых данных.

Для проведения экспериментальных исследований используется материал корпусов данных для русского языка: TeCoRus [19], RuSpeech [20] и PronExRu [21].

Данные разделены на три части: обучающую, настроечную и тестовую выборки. Обучающая выборка, на которой оценивались параметры акустических моделей, включала материал корпусов RuSpeech и TeCoRus (обучающую выборку), в основном, читаемую речь от 200+ чел. Настроечная выборка, использованная для оценки параметра λ, состояла из тестового материала корпуса TeCoRus (1000 предложений от 10 чел.). Тестовая включает материал корпуса PronExRu.

На обучающих данных были построены акустические модели звуков, контекстно-зависимые скрытые марковские модели трифонов, из трех состояний, гендер-зависимые. Всего обучались параметры около 10 000 состояний. Оценка признаков (5), за исключением темпа речи ROS, делается на основе данных из произносительного словаря, пополненного информацией о части речи слов.

Вычисление параметра ROS осуществляется на основе так называемого относительного темпа речи (relative ROS, [14]).

В ходе предварительных экспериментов для каждого центрального состояния акустических моделей рассчитывались бинарные значения признаков из (5), а по (4) вычислялись значения параметра λ интерполированных моделей.

Для ранжирования и последующего выбора значений λ по признакам вариативности строилось бинарное дерево решений. Вопросы для формирования дерева относились к наличию или отсутствию соответствующих признаков вариативности, например, «принадлежит ли звук, содержащий моделируемое состояние функциональному слову?», «принадлежит ли звук окончанию слова?» и т. п. В качестве критерия для выбора лучшей пары «вопрос–лист» дерева для разбиения на текущем шаге алгоритма использовалось изменение значения энтропии параметра λ в результате расщепления листа.

Полученные результаты имеют предварительный характер. Можно утверждать, что оптимальная величина параметра λ действительно существенно зависит от характеристик речи: относительные значения λ для читаемого материала настроечной выборки в [20] в среднем на треть меньше, чем для спонтанной речи из [21]. Более того, для обучающих данных среднее значение параметра λ оказывается ненулевым (0,12), как можно было бы ожидать. Для одного и того же типа материала наиболее существенное увеличение λ наблюдается для признаков окончания слов и функциональных слов (предлогах). Использование интерполированных моделей даже с усредненными (без классификации деревом решений) значениями λ приводит к повышению оценок правдоподобия данных для корректных гипотез, поэтому можно ожидать также снижения уровня ошибок распознавания при использовании сглаженных акустических моделей.

4 Заключение

Статья посвящена исследованию возможности снижения уровня ошибок при автоматическом распознавании естественной речи за счет использования неявных моделей вариативности произношения. В качестве основного источника вариативности рассматривается нечеткая, неполная артикуляция как следствие нарушения синхронизации работы частей речеобразующего тракта. Для учета такого типа вариативности предложено заменить исходные акустические модели их комбинациями в виде сглаживания параметров текущих моделей параметрами последующих. В ходе предварительных численных экспериментов на корпусах данных показано, что вариативность произнесения корректно рассматривать как временный фактор, обусловленный текущим фонетическим, позиционным и просодическим контекстом, соответственно параметры сглаживания также должны быть контекстно-зависимыми. Приведены контекстные признаки появления вариативности и показано, что использование сглаженных моделей звуков в потенциально вариативных позициях повышает правдоподобие данных, что коррелирует со снижением уровня ошибок при распознавании разговорной речи.

Литература

[1] Word error rate. `http://en.wikipedia.org/wiki/Word_error_rate`.

[2] *McAllaster D., Gillick L., Scattone F., Newman M.* Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch // Conference (International) on Speech and Language Processing. — Sydney, 1998. P. 1847–1850.

[3] *Saraclar M., Nock H., Khudanpur S.* Pronunciation modeling by sharing Gaussian densities across phonetic models // Comput. Speech Lang., 2000. Vol. 14. No. 4. P. 137–160.

[4] *Wester M.* Pronunciation modeling for ASR — knowledge-based and data-derived methods // Comput. Speech Lang., 2003. Vol. 17. P. 69–85.

[5] *Saraclar M., Khudanpur S.* Pronunciation change in conversational speech and its implications for automatic speech recognition // Comput. Speech Lang., 2004. Vol. 18(4). P. 375–395.

[6] *Jelinek F.* Statistical methods for speech recognition. — Cambridge, MA, USA: MIT Press, 1997. 305 p.

[7] *Lehr M., Gorman K., Shafran I.* Discriminative pronunciation modeling for dialectal speech recognition // International Speech Communication Association, Interspeech Conference Proceedings. Singapoure, 2014. P. 1458–1462.

[8] *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. — USA, 1997. P. 26–33. doi: 10.1109/ASRU.1997.659004.

[9] *Hitchinson B., Droppo J.* Learning non-parametric models of pronunciation in automatic speech recognition // Conference (International) on Acoustics, Speech, and Signal Processing Proceedings. — USA, 2011. P. 4904–4907.

[10] *Schramm H.* Modeling spontaneous speech variability for large vocabulary continuous speech recognition. Germany: Technical University of Aachen, 2006. D.Sc. Diss.

[11] *Livescu L., Glass J.* Feature-based pronunciation modeling for speech recognition // Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics Proceedings. — New York, NY, USA, 2004.

[12] *Hain T, Woodland P. C.* Dynamic HMM selection for continuous speech recognition // Proc. EuroSpeech, 1999. P. 1327–1330.

[13] *Hain T.* Implicit modelling of pronunciation variation in automatic speech recognition // Speech Commun., 2005. Vol. 46. P. 171–188.

[14] *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.

[15] *Liu Y.* Modeling partial pronunciation variations for spontaneous Mandarin speech recognition // Comput. Speech Lang., 2003. Vol. 17. No. 4. P. 357–379.

[16] *Spiess T., Wrede B., Fink G. A., Kummert F.* Data-driven pronunciation modeling for ASR using acoustic subword units // Conference (International) on InterSpeech, 2003. P. 2549–2552.

[17] *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.

[18] *Rabiner L., Biing-Hwang J.* Fundamentals of speech recognition. — Signal processing ser. — New Jersey, USA: Prentice Hall, 1993. 496 p.

[19] *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.

[20] *Bogdanov D. S., Krivnova O. F., Podrabinovitch A. J., Arlazarov V. L.* Creation of Russian Speech Databases: Design, processing, development tools // Conference (International) on Speech and Computers Proceedings. Moscow, 2004. С. 650–656.

[21] База фрагментов разговорной русской речи. Свидетельство о регистрации базы данных 2016620687, 2016.

Поступила в редакцию 01.09.2016

Implicit pronunciation variation model for automatic speech recognition*

V. J. Chuchupal
v.chuchupal@gmail.com
Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

The variations in pronunciation of words in natural speech are considered as one of the main sources of speech recognition errors. This is the reason for development and implementation of the advanced pronunciation models in modern ASR (automatic speech recognition) systems. The paper considers the pronunciation variations that are caused by a fuzzy or an incomplete articulation that is frequently observed in spontaneous speech. The author proposes the use of the implicit pronunciation model that is implemented as the combination of the acoustical models of the adjacent phones. Such a model could be realized by smoothing or interpolation of the corresponding model parameters. Also, it is proposed to use the context-dependent interpolation, so that the values of the smoothing parameters are conditioned by the current position, syntax, and prosodic contexts of the sound. While the pronunciation modeling approach on the base of combination of acoustical models (including the interpolation) has already been discussed in literature, the proposed method based on the combination of the adjacent models with the use of the context-dependent smoothing parameters has not already been published as far as the author knows. The numerical experiments on the databases that contained both the read and spontaneous speech showed the correctness of the proposal about the use of the acoustic model combination on the base of interpolation and proposal to utilize the variable smoothing parameters such that the parameter values are conditioned on the features of phonemic context, syntax, and prosody.

Keywords: *automatic speech recognition; pronunciation variation; pronunciation modeling; hidden markov models*

DOI: 10.21469/22233792.2.4.01

References

[1] Word error rate. Available at: `http://en.wikipedia.org/wiki/Word_error_rate` (accessed January 10, 2017).

[2] McAllaster, D., L. Gillick, F. Scattone, and M. Newman. 1988. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. *Conference (International) on Speech and Language Processing*. Sydney. 1847–1850.

[3] Saraclar, M., H. Nock, and S. Khudanpur. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Comput. Speech Lang.* 14(4):137–160.

[4] Wester, M. 2003. Pronunciation modeling for ASR — knowledge-based and data-derived methods. *Comput. Speech Lang.* 17:69–85.

[5] Saraclar, M., and S. Khudanpur. 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Comput. Speech Lang.* 18(4):375–395.

[6] Jelinek, F. *Statistical methods for speech recognition.* Cambridge, MA: MIT Press, 1997. 305 p.

[7] Lehr, M., K. Gorman, and I. Shafran. 2014. Discriminative pronunciation modeling for dialectal speech recognition. *International Speech Communication Association, Interspeech Conference Proceedings*. Singapoure. 1458–1462.

*The research was supported by the Russian Foundation for Basic Research (grant 14-01-00607).

[8] Byrne, B., M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos. 1997. Pronunciation modelling for conversational speech recognition: A status report from WS97. *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. 26–33. doi: 10.1109/ASRU.1997.659004.

[9] Hitchinson, B., and J. Droppo. 2011. Learning non-parametric models of pronunciation in automatic speech recognition. *Conference (International) on Acoustics, Speech, and Signal Processing Proceedings*. USA. 4904–4907.

[10] Schramm, H. Modeling spontaneous speech variability for large vocabulary continuous speech recognition. Gernany: Technical University of Aachen. D.Sc. Diss.

[11] Livescu, L., and J. Glass. 2004. Feature-based pronunciation modeling for speech recognition. *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics Proceedings*. New York, NY.

[12] Hain, T., and P. C. Woodland. 1999. Dynamic HMM selection for continuous speech recognition. *Proc. EuroSpeech*. 1327–1330.

[13] Hain, T. 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Commun.* 46:171–188.

[14] Zheng, J., H. Franco, and A. Stolcke. 2003. Modeling word-level rate-of-speech variation in large vocabulary conversational speech recognition. *Speech Commun.* 41:273–285.

[15] Liu, Y. 2003. Modeling partial pronunciation variations for spontaneous Mandarin speech recognition. *Comput. Speech Lang.* 17(4):357–379.

[16] Spiess, T., B. Wrede, G. A. Fink, and F. Kummert. 2003. Data-driven pronunciation modeling for ASR using acoustic subword units. *Conference (International) on InterSpeech*. 2549–2552.

[17] Ostendorf, M., I. Shafran, and R. Bates. 2003. Prosody models for conversational speech recognition. *2nd Plenary Meeting and Symposium on Prosody and Speech Processing*. USA. 147–154.

[18] Rabiner, L., and J. Biing-Hwang. 1993. *Fundamentals of speech recognition*. Signal processing ser. New Jersey: Prentice Hall. 496 p.

[19] Chuchupal, V. J., K. A. Makovkin, A. V. Chichagov, V. B. Kuznetsov, and V. F. Ogaryshev. 2005. Speech corpus TeCoRus. The sertificate of database registration No. 2005620205. RosPatent.

[20] Bogdanov, D. S., O. F. Krivnova, A. J. Podrabinovitch, and V. L. Arlazarov. 2004. Creation of Russian Speech Databases: Design, processing, development tools. *Conference (International) on Speech and Computers Proceedings*. Moscow. 650–656.

[21] Corpus of spontaneous speech in Russian. 2016. The sertificate of database registration No. 2016620687. RosPatent.

Received September 1, 2016