О возможности идентификации говорящего с использованием Skyp-канала (на базе акустических параметров)*

Потапова Р.К., *академик Международной академии информатизации, доктор филологических наук, профессор, заслуженный работник Высшей школы РФ*

Собакин А.Н., *доктор филологических наук,*

Маслов А.В., *преподаватель*

В статье предложен метод идентификации говорящего по речевому сигналу в системе Skype на базе импульсного преобразования речи (ИПР). Для сравнения исследовались речевые сигналы, записанные в безэховой камере, и те же речевые сигналы, прошедшие через канал передачи IP-телефонии Skype. Цель исследования — определение индивидуальных особенностей функционирования голосового источника говорящего (фонации) в зависимости от канала передачи речевого сигнала для установления возможности идентификации говорящего по голосовым характеристикам в информационных системах.

• метод ИПР • характеристики голосового источника • статистическое оценивание формы импульса • интрадикторская вариативность.

The paper presents the research method proposed for speaker identification by speech signals in the Skype system based on the use of speech pulse conversion (SPC). Speech signals recorded in an anechoic chamber, and the same speech signals transmitted via the IP-telephony Skype channel were studied. The goal of the research was to identify individual features of speaker's vocal source (phonations) depending on the speech signal transmission channel in order to ascertain the possibility of speaker identification by speech characteristics in information systems.

---

* Доклад был прочитан на «XIV Международной научно-технической конференции. Киберненика и высокие технологии XXI века»

Идентификация говорящего по голосу и речи, пути и методы решения этой задачи становятся всё более актуальными аспектами проблемы выявления лиц, выступающих в деструктивном ключе и призывающих к дестабилизации государственных устоев, терроризму, смене режимов и т.д. Следует подчеркнуть, что развитие новых информационн0-коммуникационных технологий и, в частности, Интернета внесло свой «вклад» в социальную составляющую межличностного взаимодействия.

В настоящее время в Интернете активно используется наряду с письменной речью звучащая (устная) речь на базе программы Skype. Согласно данным средств массовой информации (например, Metro, № 11, 02.2013 г.) аналитики компании TeleGeography отмечают, что в настоящее время интернет-сервис Skype занимает одну треть от всего телефонного голосового трафика в мире. Если по обычной телефонной связи за 2012 год объём передаваемой информации увеличился на 5% по сравнению с показателем 2011 года, то по связи Skype — на 44%, что свидетельствует о возросшей роли данного способа коммуникации.

В связи с этим возникает новая опасность использования данного инструментария в деструктивных целях. Вполне вероятно, что использование методов маскировки (например, грима, накладных бороды, усов и т.д.) может изменить визуальный образ говорящего по каналу Skype. Однако можно предположить, что определенные параметры речевого сигнала говорящего, несмотря на влияние передающего тракта, несут «следы» индивидуального голоса и речи говорящего, участвующего в акте речевой коммуникации, и это может относиться, прежде всего, к тонкой структуре речевого сигнала.

Процесс речевой коммуникации представляет собой сложный и не полностью изученный феномен [Потапова 2010; 2012]. Одним из подходов к изучению подобного феномена является процедура упрощения явления, выделения его наиболее важных характеристик и функциональных связей. Оптимальным подходом к исследованию речевой коммуникации служит создание функциональной модели процесса речевой коммуникации, включающего блоки речепроизводства и речевосприятия.

Процесс речепроизводства можно представить в виде двух компонент:

— формирование команд управления на нейронном уровне органами фонации и артикуляции;

— непосредственное генерирование органами фонации и артикуляции звуковых эффектов, соотносящихся с частотой основного тона и сегментным строем того или иного языка.

Органы слуха реципиента регистрируют волновые колебания воздушного давления, и высшие отделы головного мозга производят дальнейшую обработку данных. Все перечисленные уровни речевой коммуникации труднодоступны для прямых наблюдений и регистрации характеристик их функционирования, что создает дополнительные трудности при изучении данного явления.

В ряде прикладных областей речеведения [Потапова 2010; Потапова, Потапов 2012] единственным доступным для измерений является речевой сигнал, который служит в дальнейшем базовым источником информации о процессах речеобразования и восприятия речи, их параметрах и характеристиках.

Опираясь на исследования пространственного распространения звукового давления в речевом тракте, Г. Фант обобщил полученные результаты [Фант 1964] и предложил одномерную модель речеобразования, удобную для разработки математических методов анализа и синтеза речи по параметрам. Согласно данной модели, речь рассматривается в виде фильтрации источников звука линейной системой речевого тракта. Модель является электротехническим аналогом линейной цепи с сосредоточенными параметрами в виде четырехполюсника, на входе которого имеется энергетический источник напряжения. Передаточная функция четырёхполюсника описывает резонансные свойства речевого тракта [Фант 1964: 39–58].

Источник напряжения имитирует работу голосовых связок на озвученных участках речи и (или) стохастическую функцию возбуждения на шумоподобных участках. На основе линейной модели речеобразования разработан метод импульсного преобразования речи (ИПР), позволяющий по речевому сигналу исследовать характеристики функционирования голосовых связок. Исследование формы импульса основного тона (ОТ) предлагается осуществлять на основе импульсного преобразования речи (ИПР) [Собакин 1972; 2006], которое позволяет по речевым колебаниям определять аналог импульса ОТ без использования дополнительных каналов измерения акустических характеристик речевых колебаний. Метод ИПР показал свою эффективность при исследовании колебаний голосовых связок по речевым сигналам, записанным в акустической студии.

В настоящей статье рассматривается проблема применимости данного метода при исследовании сигналов, преобразованных в системах IP-телефонии как наиболее распространённых коммуникационных средств в Интернете.

В качестве примера синтетической передачи речевых сообщений в интернет-пространстве выбрана система Skype. Выбор системы передачи речи типа Skype определяется широким охватом интернет-пользователей, что может быть применено к сфере идентификации автора сообщения в специальных целях (например, для выявления авторства высказываний).

Основные задачи, связанные с применением ИПР к исследованию голосового источника и возможной идентификации говорящего по полученным данным, состоят в следующем:

1. Проверить работоспособность метода ИПР на речевых сигналах, записанных в системе Skype.

2. Провести сравнительный анализ полученных данных с результатами исследования тех же речевых сигналов, записанных в акустической студии (безэховой камере).

3. Установить наличие или отсутствие индивидуальных характеристик голосовых источников дикторов, обусловленных спецификой передачи по каналу Skype.

Необходимость проверки работоспособности метода ИПР объясняется скрытостью принципов кодирования, передачи и синтеза речевых сообщений в системе Skype. Каждая фирма, разрабатывающая систему анализа и синтеза речи (и не только), в современных условиях при наличии высокопроизводительных вычислительных систем в качестве конечного результата имеет программный продукт. Это позволяет не раскрывать методов и алгоритмов преобразования речи на этапе его первичного описания, методов кодирования и передачи по каналу связи, а также методов синтеза речевых сообщений.

В рамках данного исследования наибольший интерес представляют методы восстановления (имитации) формы импульса основного тона. При этом остаются неизвестными ни методы синтеза импульсов ОТ, ни характеристики голосового источника, используемые при синтезе голосового возбуждения речевого тракта.

При изучении речи рассматривают четыре основных типа источника речевых колебаний [Сапожков 1963, 31–37]:

1) голосовое (тональное) возбуждение;

2) шумовое (турбулентное) возбуждение;

3) смешанное возбуждение;

4) импульсное возбуждение (взрыв).

Различают также аспирированные (придыхательные) звуки и звуки с модуляцией воздушного потока, когда имеется сужение в голосовой щели, но колебаний связок не происходит.

Звуки речи делятся на гласные (только голосовой источник возбуждения), глухие согласные (турбулентный источник возбуждения), звонкие согласные (смешанный источник возбуждения) и взрывные. Тип источника возбуждения является одним из признаков звука и его отличительной характеристикой. Наиболее информативным из перечисленных источников является голосовой источник возбуждения речевого тракта [Ondrachkova 1966; Потапова, Михайлов 2012; Потапова, Потапов 2012]. Колебания голосовых связок в процессе образования звуков речи содержат индивидуальные особенности говорящего и улавливаются слушающим в процессе восприятия слухового образа. Особенно ярко индивидуальные характеристики гортани проявляются при образовании гласных звуков речи.

Гласные звуки речи встречаются в различном окружении других звуков речи и, в частности, могут образовываться в изолированном (стационарном) варианте. При этом функционирует только голосовой источник возбуждения.

Основными характеристиками голосового источника являются:

1) интенсивность голосового источника — временна́я огибающая речевого сигнала;

2) период колебаний голосовых связок (период основного тона) или величина обратная этому периоду — частота основного тона;

3) форма импульса основного тона, определяемая микроструктурой колебаний голосовых связок в процессе образования гласных звуков речи.

Первые два признака голосового источника (интенсивность и частота основного тона) вместе с формантными характеристиками используются во многих системах идентификации дикторов [Рамишвили 1981] и, как отмечает автор, значительная часть информации о тембре голоса диктора при таком подходе не учитывается. Это обстоятельство оказывает отрицательное влияние на качество процедуры идентификации дикторов.

Таким образом, представляется целесообразным при идентификации использовать информацию о форме импульса основного тона или о его аналоге. Задача определения формы импульса основного тона по речевому сигналу является достаточно сложной и до настоящего времени не имеющей эффективного решения. Один из возможных вариантов ее решения описан в работе [Собакин 2010].

Форма импульса основного тона в модельном представлении иногда описывается положением максимума импульса, величиной (амплитудой) максимума, наклоном начального и конечного участков импульса [Fant 1979]. Общепринято, что максимум возбуждения речевого тракта определяется крутизной наклона импульса глотки в момент закрывания связок. Однако более тонкие исследования функционирования голосового источника показывают, что аппроксимация импульса основного тона прямыми линиями и его экстремумами не учитывает продольных и поперечных колебаний голосовых связок.

По данным работы [Sundberg, Gauffin 1978] на интервале смыкания голосовых связок воздушный поток уменьшается до нуля не сразу, а постепенно. Это объясняется тем, что голосовые связки расположены в процессе сближения не параллельно. Иногда на интервале смыкания отмечается небольшая пульсация.

В данном исследовании используется импульсное преобразование речи (ИПР) [Собакин 1972; 1999; 2006] для выявления индивидуальных характеристик голосовых связок по речевому сигналу.

Преобразование речи в импульсную последовательность, синхронную с колебаниями голосовых связок, позволяет исследовать форму полученных импульсов методами математи-

ческой статистики. Для этого предлагается проводить нормировку полученных импульсов по их центрам, и осуществлять сложение нормированных импульсов. Эти процедуры позволяют получить статистически значимый «образ» полученной последовательности импульсов в виде нечеткого множества, сохраняющий индивидуальные особенности функционирования голосового источника диктора.

В ходе исследования аппаратно-программный комплекс, при помощи которого осуществлялась звукозапись в акустической студии, состоял из следующих элементов:

I. *Аппаратное обеспечение*

I.1. Портативный компьютер Toshiba Qosmio G30-211-RU

I.2. Аудиоплата Creative Professional E-MU 1616m PCMCIA

I.3. Усилитель измерительный Bruel & Kjaer, тип 2610

I.4. Микрофон Shure SM48, класс качества 1

I.5. Дикторская кабина AUDIOSTOP-1

II. *Программное обеспечение*

II.1. Операционная система Microsoft Windows XP

II.2. Программа записи и исследования речевых сигналов «Мастерская сигналов» («ZSignalWorkshop» [Мастерская сигналов 2012]).

II.3. Виртуальный микшерский пульт Creative (поставляется в комплекте со звуковой платой, указанной в п. I.2).

В целях мониторинга и контроля качества производимых записей дополнительно использовались наушники профессионального класса качества Beyerdynamic DT 770 PRO, программно-аппаратный комплекс Computerized Speech Lab (CSL), модель 4500, а также программный пакет Sony Sound Forge 9.0.

В рамках данного исследования проверялась гипотеза о возможности определения индивидуальных характеристик импульса основного тона по речевому сигналу с использованием ИПР для двух типов записи речевых сигналов:

1) *сигналов, записанных в акустической студии и пропущенных через полосовой фильтр телефонного канала связи (сигналы без искажений системы передачи по каналу связи);*

2) *сигналов первого типа, пропущенных через систему телефонии Skype*.

Обработка гласных осуществлялась вычислительной программой в системе символьной математики MATLAB 7.6.0.324, автором которой является А.В. Маслов.

На первом этапе исследовались сигналы первого типа без искажений канала связи.

Выбранный исследователем отрезок речевого сигнала (рис. 1) делится на вектора (одномерные массивы) типа $\{x(j),\ldots, x(j+N+p-1)\}$,где N- размерность векторов, p — порядок автокорреляционной матрицы. Параметры задаются исследователем.

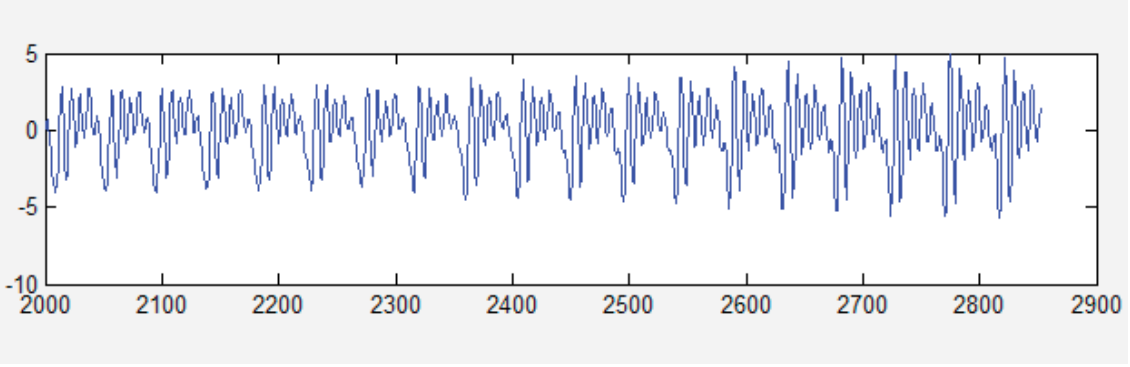

Рис. 1. Осциллограмма звука [*a*] диктора — женщины

По набору векторов строится соответствующая автокорреляционная матрица. Последовательно с изменением текущего значения временно́го параметра *n* вычисляется определитель каждой построенной матрицы. Для сглаживания выделенных импульсов в некоторых случаях применялся корень из определителя порядка автокорреляционной матрицы.

Последовательность вычисленных определителей образует импульсную функцию, являющуюся моделью работы голосовых складок (рис. 2).

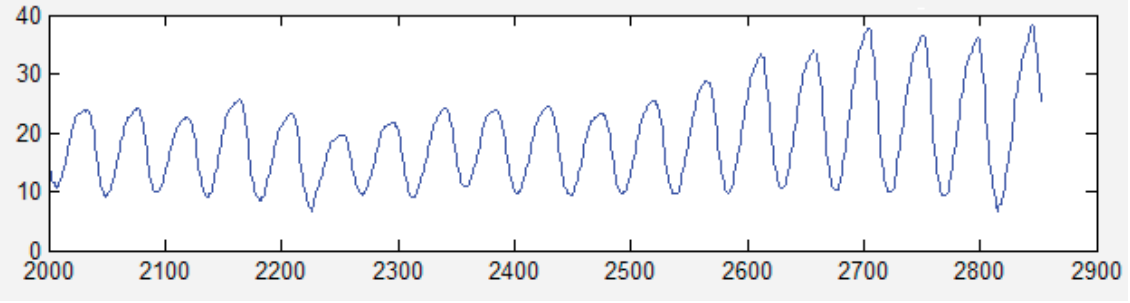

Рис. 2. Последовательность импульсов

Полученная квазипериодическая последовательность импульсов полностью согласуется, по крайней мере визуально, с работой голосового источника: импульсы соответствуют увеличению амплитуды речевых колебаний на периоде основного тона. Именно так в рамках рассматриваемой модели речеобразования должно происходить в моменты раскрытия голосовых связок: проступающая в речевой тракт (линейную диссипативную систему) энергия подсвязочного давления увеличивает амплитуду сигнала на выходе этой системы.

Дальнейшая статистическая обработка полученных импульсов направлена на получение статистически значимой оценки формы («образа») импульса основного тона. Подобная отработка полученной импульсной последовательности связана с решением нескольких сложных задач:

1) выделение самих импульсов;
2) нормировку импульсов по амплитуде;
3) статистическое оценивание формы импульса.

В импульсной последовательности определялись экстремумы (максимумы и минимумы соответственно), что позволяло выделить изолированные импульсы. При этом значения последовательности, не превышающие пороговой величины, обнулялись.

В данной работе пороговое значение принималось равным 1/100 от максимума одного импульса или временной огибающей, вычисленной для нескольких (как правило, 3-5) смежных импульсов.

Пример описанного метода выделения импульсов приведен на рисунке 3.

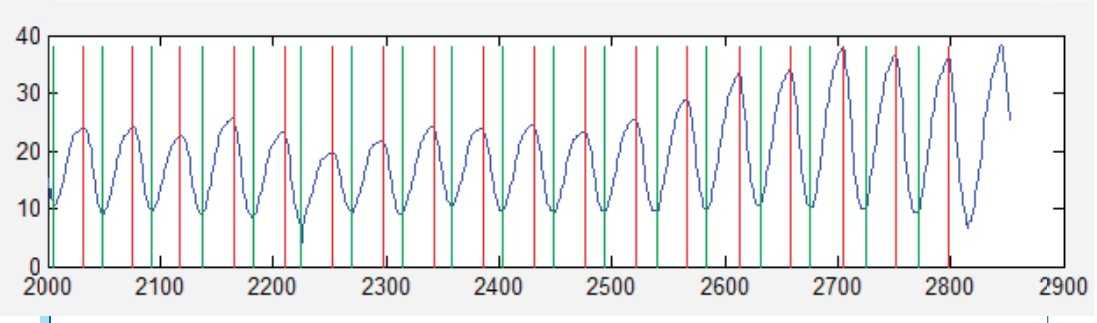

Рис. 3. Выделенные максимумы и минимумы

Изолированный импульс в данном случае определялся как часть изменения функции на участке от минимума до ближайшего минимума. Из приведённого примера видно, что полученные таким образом импульсы будут отличаться друг от друга по амплитуде и длительности. Поэтому простое наложение этих импульсов друг на друга является некорректным и подобная процедура не позволяет в данном случае выделить статистически значимую форму импульса основного тона.

В данной работе предлагается использовать интегральные характеристики выделенного импульса, зависящие от всей конфигурации импульса в целом. Подобной интегральной характеристикой импульса, по мнению авторов проекта, может служить центр импульса, координаты которого определяются равенством площадей относительно двух секущих параллельных осям координат.

Полученные результаты содержат информацию о двух принципиально важных свойствах работы голосовых связок в процессе образования изолированных гласных звуков речи:

— характер колебаний голосовых связок коррелирован с произносимым звуком речи для одного и того же диктора;

— голосовые связки разных дикторов при образовании одинаковых звуков функционируют по-разному и имеют несомненные индивидуальные характеристики.

Напомним, что в данной работе рассматривается вопрос о применимости импульсного преобразования речи к исследованию сигналов, преобразованных в системах IP-телефонии как наиболее распространенных коммуникационных средств интернета.

Как указывалось ранее, в качестве примера синтетической передачи речевых сообщений в интернет-пространстве была выбрана распространенная система Skype, т.е. рассматривались сигналы второго типа, прошедшие си-

стему Skype, и содержащие дополнительные шумы и искажения, свойственные данной системе передачи речевых сообщений.

В рамках настоящего исследовании наибольший интерес представляют методы восстановления (имитации) формы импульса основного тона в системе Skype. При этом остаются неизвестными ни методы синтеза импульсов ОТ, ни характеристики голосового источника, используемые при синтезе голосового возбуждения речевого тракта.

Работоспособность метода проверялась на материале, полученном при произнесении шести дикторов. Каждый диктор произносил в стационарном режиме шесть русских гласных «а», «э», «и», «о», «у», «ы». Гласные записывались в акустической комнате (безэховой камере) с помощью широкополосного микрофона. Записанные гласные пропускались через систему передачи речевых сообщений Skype и вновь записывались. Качество записи исходных и преобразованных гласных в обоих случаях было высоким (частота дискретизации по времени — более 40 кГц, дискретизация по амплитуде — 16 бит/отсчет).

Записанные сигналы гласных программно фильтровались полосовым фильтром с полосой пропускания от 300 Гц до 3400Гц с затуханием порядка 60 дБ на концах диапазона частот от 0Гц до 4000Гц, соответствующего диапазону частот телефонного канала. Затем сигнал прореживался в отношении 1:5, что приближенно соответствовало частоте дискретизации по времени 8 кГц (по Котельникову).

Предварительный ответ на вопрос, поставленный в ходе данного исследования, следует считать положительным. Метод ИПР работоспособен в применении к сигналам в системе Skype при исследовании индивидуальных характеристик голосового источника, как для мужских, так и женских голосов.

Результаты исследования изолированно произнесенных русских гласных на данном этапе позволяют сделать следующие выводы:

1. Метод ИПР работоспособен на речевых сигналах (гласных звуках речи), записанных на приёмном конце IP-телефонии системы Skype.

2. Сравнительный анализ импульсных характеристик гласных, полученных в системе Skype, с результатами тех же речевых сигналов, записанных в акустической студии (безэховой камере), показал, что усредненные «образы» импульсов отличаются друг от друга для одного и того же диктора, т.е. имеется интрадикторская вариативность.

3. Выделенные импульсные последовательности и созданные на их основе статистически значимые «образы» импульсов основного тона в системе Skype, описывающие работу голосового источника, по форме различны для различных гласных одного диктора и, что более важно, ***существенно отличаются друг от друга для разных дикторов (интердикторская вариативность)***, что может служить основой для идентификации говорящего.

Литература

1. *Потапова Р.К*. Речь: Коммуникация, информация, кибернетика. 4-е изд., доп. М.: Книжный дом «Либроком», 2010. 594 с.

2. *Потапова Р.К*. Речевое управление роботом: лингвистика и современные автоматизированные системы. 3-е изд. М.: Комкнига, 2012. 328 с.

3. *Потапова Р.К*. Новые информационные технологии и лингвистика: Учебное пособие. Изд. 5-е. М.: Книжный дом «ЛИБРОКОМ», 2012. 368 с.

4. *Потапова Р.К., Михайлов В. Г*. Основы речевой акустики. М.: ИПК МГЛУ «Рема», 2012. 494 с.

5. *Потапова Р. К., Потапов В. В*. Речевая коммуникация: От звука к высказыванию. М.: ЯСК, 2012. 464 с.

6. *Рамишвили Г.С*. Автоматическое опознавание говорящего по голосу. М.: «Радио и связь», 1981. 224 с.

7. *Сапожков М.А*. Речевой сигнал в кибернетике и связи. М.: «Связьиздат», 1963. 452 с.

8. *Собакин А.Н*. Об определении формантных параметров голосового тракта по речевому сигналу с помощью ЭВМ // Акустический журнал АН СССР. 1972. № 1. С. 106–114.

9. *Собакин А.Н*. Основной тон речи и метод его исследования // IX сессия РАО: Современные речевые технологии. Сб. тр. М.: ГЕОС. 1999. С. 47–50.

10. *Собакин А.Н*. Артикуляционные параметры речи и математические методы их исследования // Вестник МГЛУ. М., 2006. 220 с.

11. *Собакин А.Н*. Выделение импульсов основного тона по речевому сигналу. XXII сессия РАО: Современные речевые технологии. Сб. тр. М.: ГЕОС. 2010. С. 48–52.

12. *Фант Г*. Акустическая теория речеобразования / Пер. с англ. М.: Наука, 1968. 284 с.

13. *Fant G*. Speech production. Glottal source and excitation analysis // Quart Progr. And Status. Rept. Speech Transmiss. Lab. 1979. № 1. P. 85–107.

14. Farnsworth D.W. High speed motion pictures of the human vocal cords // Bell Teleph. Lab.: Record. 1940. V. 18. P. 203.

15. *Ondrachkova J*. Glottographical research in sound groups // Модели восприятия речи. Международный психологический конгресс. М.,1966. Л., 1966. P. 90-94.

16. *Sundberg J. and Gauffin J*. Logopedics wave-form and Status Rept. Speech Transmiss. Lab. 1978. № 2–3. P. 35–50.

17. Электронный ресурс «Мастерская сигналов», 2012. Режим доступа [http://zhenilo.narod.ru/main/index.htm].

Сведения об авторах

Потапова Родмонга Кондратьевна —
доктор филологических наук. Академик Международной академии информатизации, профессор, директор Института прикладной и математической лингвистики ф-та ГПН МГЛУ, заслуженный работник высшей школы Российской Федерации,

Собакин Аркадий Николаевич —
кандидат технических, доктор филологических наук, ФГБОУ ВПО Московский государственный лингвистический университет,

Маслов Алексей Витальевич —
преподаватель кафедры прикладной и экспериментальной лингвистики ф-та ГПН МГЛУ.