ПРИМЕНЕНИЕ МЕТОДА ФОРМАНТНОГО АНАЛИЗА ДЛЯ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ АКУСТИЧЕСКОГО СИГНАЛА В ЗАДАЧАХ РАСПОЗНАВАНИЯ РЕЧИ

Герасимов А.В.

Научно-исследовательский физико-технический институт нижегородского государственного университета им. Н.И. Лобачевского (НИФТИ ННГУ)
603600, Россия, Н. Новгород, пр. Гагарина 23, корпус 3, estro@nifti.unn.ru

ВВЕДЕНИЕ

Для описания процесса распознавания речи было предложено достаточно большое количество различных алгоритмов, но ни один из них не позволяет достичь человеческого уровня распознавания. Алгоритмы извлечения смысловой значащей информации из речевого сигнала опираются на предположения/исследования относительно модели генерации сигнала, пытаясь максимально учесть особенности образования речи и ее восприятия человеком[1]. Целью данной работы является исследование алгоритма «анализа через синтез» на основе линейного предсказания.

МОДЕЛЬ "АНАЛИЗА ЧЕРЕЗ СИНТЕЗ"

Существует несколько способов моделирования (синтеза) речевого сигнала. Наиболее адекватной реальному голосовому аппарату является линейная модель, относящаяся к группе параметрических моделей синтеза речевого сигнала, основывающаяся на его устройстве. Минуя задачи моделирования колебания связок и формирования резонансных полостей, рассматривая только изменения волнового сигнала, получаем схему, изображенную на рис.1: [2]

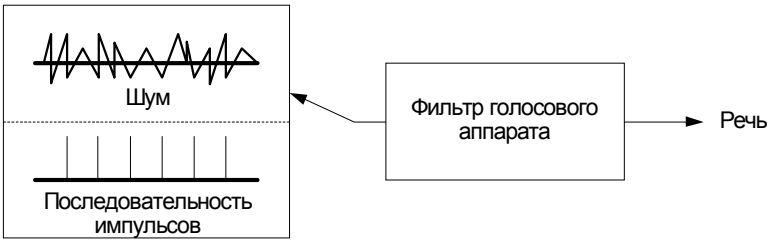

Рис. 1: Схема параметрической модели речеобразования.

В данной модели выходной сигнал представляется в виде свертки возбуждающего сигнала, генерируемого связками и модулирующего, являющегося характеристической функцией формы ротовой полости или артикуляторной характеристикой [1]. Математически это можно описать следующей формулой:

$$s(n) = v(n) \otimes h(n) \quad (1)$$

где *v(n)* – возбуждающий сигнал, *h(n)*– модулирующий,
или в терминах **z**-преобразования:

$$S(z) = V(z) * H(z) \quad (2)$$

Возбуждающий сигнал характеризуется так называемой «просодической информацией», т.е. высотой и тембровой окраской. Эта информация может быть использована в задачах идентификации говорящего по голосу. Модулирующий сигнал рассматривается как характеристика формирующего звуки голосового тракта человека и применяется в задачах распознавания речи.

В спектральной области операция свертки двух сигналов представляется в виде произведения их образов. Возбуждающий сигнал в рамках описанной модели представляет собой либо полигармонический сигнал (в случае гласного звука), либо широкополосный шумовой (в случае согласного). Модулирующая функция представляет собой огибающую результирующего сигнала. Таким образом, задача получения аутентичной информации (при распознавании фонем) сводится к определению огибающей мгновенного спектра сигнала или так называемому формантному анализу. Поскольку модулирующую функцию можно рассматривать как передаточную функцию линейного КИХ-фильтра, значения этой функции (коэффициенты фильтра) определяются с помощью метода линейного предсказания. Данный алгоритм широко применяется в вокодерном кодировании[4].

При формантном анализе текущую оценку отсчета сигнала определяют как сумму *P* предшествующих отсчетов.

$$\tilde{s}(n) = \sum_{k=1}^{p} s(n-k) * a_k \qquad (3)$$

где, $\boldsymbol{a}=\{a_k\}$ – вектор коэффициентов предсказания. Порядок *P* при формантном анализе выбирают равным 8-12 [4].

Разность между истинным и предсказанным значением отсчета определяет ошибку предсказания или остаточный сигнал:

$$r(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} s(n-k) * a_k \qquad (4)$$

В результате z-преобразования разностного уравнения (4) имеем

$$R(z) = S(z) * A(z) \qquad (5)$$

где функция

$$A(z) = 1 - \sum_{k=1}^{p} a_k * z^{-k} \qquad (6)$$

является передаточной характеристикой цифрового фильтра, частотная характеристика которого обратна по отношению к частотной характеристике голосового тракта (если сравнить с формулой 2). Математически сказанное иллюстрируется так:

$$A(z) = \frac{1}{H(z)} \qquad (7)$$

Значения коэффициентов a (формула 6) подбираются так, чтобы минимизировать среднеквадратичное значение остаточного сигнала r. Полученные коэффициенты фильтра a можно рассматривать как вектор признаков фонемы. Для проверки степени стабильности и инвариантности получаемого вектора признаков a необходимо исследовать предел его изменений в условиях различного произношения опорной фразы (артикуляторные характеристики которой предполагаются стабильными). Вариации произношения обеспечиваются различной высотой произношения (pitch).

На рис. 2 в левом ряду представлены усредненные спектральные характеристики фонемы "a", произнесенной одним человеком с различной высотой. Усреднение проводилось по всей выборке сигнала (10 звуковых файлов, каждый длительности приблизительно 0.25 секунды). Длина фрейма равна 1024 отсчета, частота дискретизации выбрана 44100 Гц. В правом ряду показаны основные спектральные пики (форманты) на протяжении всей длительности сигнала.

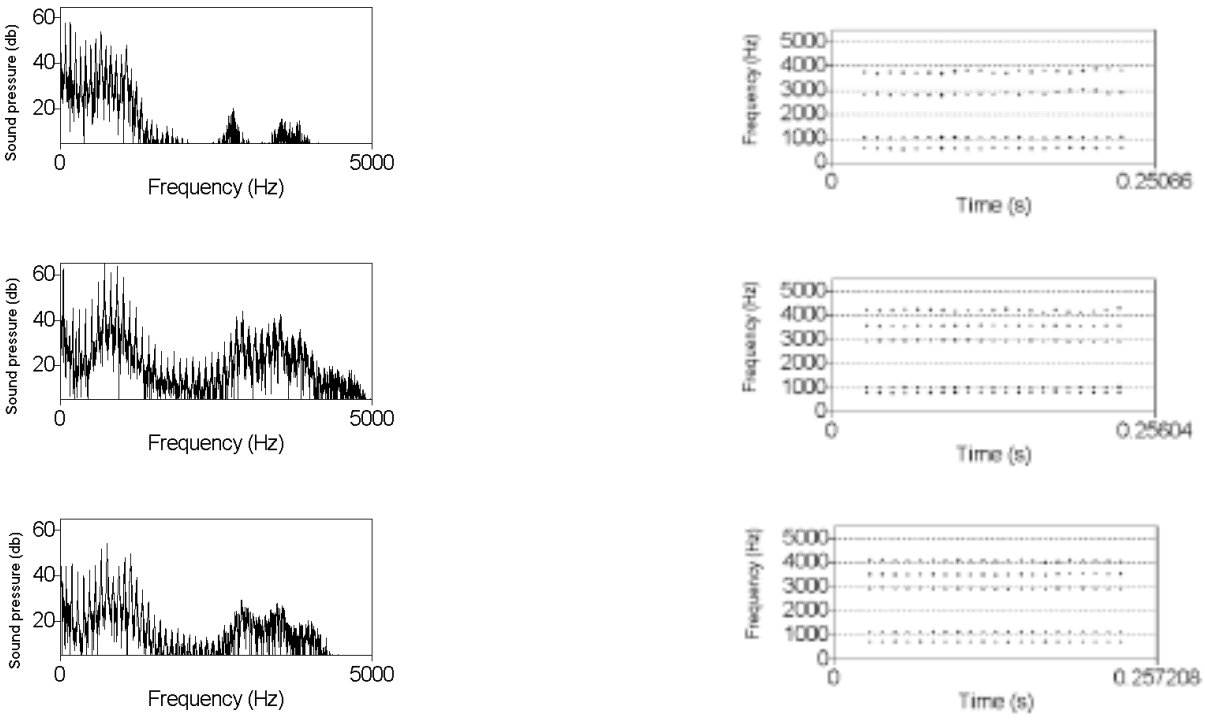

Рис.2: Усредненные спектральные характеристики фонемы "a" и соответствующие ему формантные пики на спектрограмме для высот (pitch) 650, 700 и 750 Гц соответственно.

Предполагается, что полученные данные о формантах, как и соответствующий энергетический спектр огибающей анализируемого сигнала, являются непосредственной аутентичной характеристикой фонемы. Для принятия этого предположения достаточно, доказательства того, что разброс получаемых векторов признаков в признаковом пространстве в зависимости от характеристик голоса диктора был существенно меньше, чем при различиях в зависимости от конкретных речевых единиц. Тогда полученные признаки можно будет считать инвариантными относительно высоты, тембра и громкости голоса. Инвариантность по громкости обеспечивается нормированием энергетического спектра оцениваемого сигнала.

Таким образом, определение степени инвариантности признаков сводится к исследованию смещения формантных пиков в частотной области в зависимости от диктора и произносимой. Исследования проводились для 3-х дикторов с различной высотой голоса.

Результаты представлены на диаграмме рис. 3. Абсолютное рассогласование частот соответствующих формант находится в пределах 250 герц и равно диапазону изменения высоты. Спектральный же диапазон фонемы составляет 4-5 Кгц, что говорит о незначительном изменении спектрального/формантного рисунков по сравнению с изменением частоты основного тона. Данный факт позволяет считать полученные признаки инвариантными относительно произносящего человека при условии отсутствия редукции (невнятного произношения звуков).

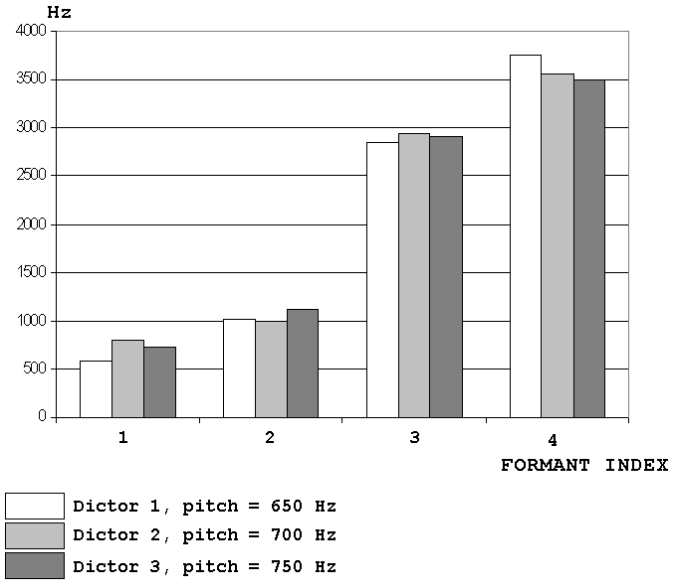

Рис. 3: Рассогласование по частотам первых 4-х формант для различных дикторов (представлены наиболее различающиеся случаи).

Очевидно, что любые неотфильтровавшиеся помехи вносят ложную информацию в определяемые признаки (артикуляторные параметры), смещая положение фонемы в пространстве признаков и понижая тем самым процент достоверности классификации. Но, как показали исследования [1], артикуляторные параметры также зависят от конкретного человека и не являются "абсолютно стабильными" речевыми характеристиками. При плохой дикции и невыразительной речи понимать слова приходится из контекста. В компьютерной модели такая обработка осуществляется на последующих этапах (фонемная категоризация, семантический анализ и т.д.). Для этапа же выделения информации о (произнесенной единице речи) описанный выше метод оказывается вполне применимым.

ЗАКЛЮЧЕНИЕ

Важным достоинством этого метода является относительная простота оценки параметров фильтра ***A(z)***, т.к. используются линейные процедуры обработки сигнала. Направлением дальнейшей работы является исследование эффективности работы алгоритма при аддитивных шумоподобных/гулоподобных искажениях в совокупности с алгоритмами предварительной коррекции, ориентированных на подавление специфических помех [3], осложняющих применение методов, основанных на выбранной модели.

ЛИТЕРАТУРА

[1] Венцов А.В., Касевич В.Б. «Современные модели восприятия речи: критический обзор», Издательство Санкт-Петербургского университета, С-Пб, 1994.

[2] Gaurang Kishor Parikh, B.E. «The effect of noise on the spectrum of speech», Thesis, Texas Un-ty, 2002

[3] Чучупал В.Я., Чичагов А.С., Маковкин К.А. «Цифровая фильтрация зашумленных речевых сигналов», ВЦ РАН, М. 1998.

[4] Маркел Дж., Грей А. Линейное предсказание речи. М.: Связь, 1980

———————♦———————

APPLICATION OF FORMANT ANALYSIS METHOD FOR FEATURE EXTRACTION FROM ACOUSTIC SIGNAL IN SPEECH RECOGNITION TASKS

Gerasimov. A.

For the speech recognition process description plenty of various algorithms was proposed, but all of them does not allow to achieve a human level of recognition. The algorithms of the semantic meaning information extraction from a speech signal are based on the assumptions / researches concerning model of generation of a signal, trying at most consider features of speech structure and its perception(recognition) by the human [1]. The purpose of the given work is the linear prediction based «analysis through synthesis» algorithm investigation.

There are some ways of modeling (synthesis) of a speech signal. The most easy and similar to the real voice device production is the linear model that can be related to parametrical speech signal synthesis models group. In the given model the target signal is represented as convolution of a stimulating signal generated by vocal chords and modulating signal. The modulating signal is regarded as the characteristic function of the mouth cavity form.

In spectral area the operation of convolution of two signals is represented as product of their spectral images. The stimulating signal within the framework of the described model represents polyharmonical signal (in case of consonant pronunciation) or wide noise (in case of fricative pronunciation). The modulating function represents bending around of a resulting signal. Thus, the task of the authentic information obtaining is reduced to definition the average of an instant spectrum of a signal or so-called formant analysis. As the modulating function can be considered as transfer function of the linear FIR-filter, the discrete series values of this function (filter coefficients) are obtained with the linear prediction method.

For received feature vector stability and invariance checking it is necessary to explore the range of its variability according to various pronunciation of a basic phrase with stable pronouncing characteristics. The variations of a pronunciation are provided with various pitch. As the investigations have shown, the absolute mismatch of appropriate formant frequencies lays within the limits of 250 hertz such the range of pitch changes. The whole phoneme spectral range makes 4-5 kHz, that allows to consider the feature vector invariant imposing conditions of reduction absence.

The important advantage of selected method is the relative simplicity of an estimation of parameters of the filter. The further work direction is the research of an overall performance of algorithm at additive noise and din distortions in aggregate with algorithms of preliminary correction.