

МЕТОДЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ В ИНСТРУМЕНТАЛЬНО-ФОНЕТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Применение статистических методов в фонетических исследованиях связано с определением закономерностей в структуре исследуемого речевого явления, поскольку звучащая речь является случайным процессом, и физические параметры звуков могут рассматриваться как случайные величины. Трактовка результатов инструментальных измерений при исследовании различных параметров звуков может считаться полной и достоверной только после соответствующего статистического анализа [1; 4; 6; 7; 8].

Основным методом, используемым при анализе данных **инструментального исследования**, является выборочный метод, центральными понятиями которого являются *генеральная совокупность* как единство всех возможных элементов, которым присущ интересующий исследователя признак, а также *выборочная совокупность* как единство части элементов, которые отбираются из генеральной совокупности для получения достоверных сведений о всей совокупности.

Для практических целей вместо полного описания статистических свойств генеральной совокупности случайных величин обычно ограничиваются только указанием некоторых частотных характеристик на базе выбирающей совокупности с помощью закона распределения. К ним относятся [8]:

1. Точка, к которой тяготеет совокупность значений случайной величины x – *среднее арифметическое значение* $\bar{x}$, определяемое по формуле:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (1)$$

2. Разброс значений x относительно среднего значения $\bar{x}$ – *дисперсия* σ^2 , определяемая по формуле:

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (2)$$

3. Поскольку дисперсия отлична от размерности исследуемой величины x , вместо нее часто применяется положительный корень из дисперсии, который называется *стандартным отклонением* или *среднеквадратическим отклонением*.

4. Отклонение кривой распределения от симметричной – *асимметрия*, определяемая по формуле:

$$A = \frac{n \cdot \sum_{i=1}^n (x_i - \bar{x})^3}{(n-1) \cdot (n-2) \cdot \sigma^3} \quad (3)$$

5. Островершинность кривой распределения – *эксцесс*, определяемый по формуле:

$$E = \frac{n \cdot (n+1) \cdot \sum_{i=1}^n (x_i - \bar{x})^4 - 3 \cdot \sum_{i=1}^n (x_i - \bar{x})^2 \cdot (n-1)}{(n-1) \cdot (n-2) \cdot (n-3) \cdot \sigma^4} \quad (4)$$

6. *Вариативность* изучаемого явления V :

$$V = \frac{\sigma}{\bar{x}} \quad (5)$$

Случайные величины распределяются относительно своего среднего согласно определенному закону распределения. Наиболее распространенным в статистических исследованиях является нормальное (гауссовское) распределение, которое может быть описано следующей функцией:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x-\bar{x}}{2\sigma^2}} \quad (6)$$

Выявление закона распределения каждой конкретной выборки является очень важным, поскольку в зависимости от этого будут использованы для ее описания те или иные статистические приемы. Различные фонетические явления могут распределяться по различным законам, отражающим природу звукового явления, и их выявление требуется в каждом конкретном случае. Если объем выборки является достаточно большим ($n > 100$), то законом распределения можно пренебречь и выборку рассматривать как нормальную, применяя к ней параметрические статистические приемы. Однако в случае выборок с меньшим объемом ($n < 100$) проверка на нормальность является обязательной, поскольку применение статистических приемов, не соответствующих данному закону распределения, может привести к ошибочным результатам.

Современные методики проведения предварительного статистического анализа массива данных рекомендуют использование W -критерия Шапиро-Уилкса, поскольку этот статистический тест не требует предварительного знания среднего значения или дисперсии гипотетической выборки, соответствующей нормальному распределению. Если полученное в результате проведения этого теста значение W существенно отличается от теоретического ($p \leq 0,05$), то можно сделать вывод о том, что данная выборка не подчиняется законам нормального распределения, и наоборот [8].

Тест Шапиро-Уилкса, однако, не дает представления о характере распределения, если определено, что оно существенно отличается от нормального. В связи с этим при проведении проверки на нормальность необходимо обращать внимание на значения асимметрии A и эксцесса E .

Наглядным способом проверки нормальности распределе-

ния массива данных является **построение гистограмм**. Гистограмма графически представляет частотное распределение избранной случайной величины x_i и строится по каждому виду сгруппированных данных. Предполагаемая область значений x_i делится независимо от выборки на некоторое количество интервалов. Над каждым интервалом строится прямоугольник, высота которого пропорциональна количеству значений x_i , входящих в данный интервал*. На рис.1 в качестве примера приведена гистограмма измерений минимальных значений интенсивности фазы смычки глухого предстыкового плозива [t].

Сравнение гистограммы с наложенным на нее графиком нормального распределения показывает их хорошее совпадение, что говорит о нормальности распределения. На рис.2, напротив, гистограмма значений мощности смычного глухого [t] значительно отличается от нормального распределения, представ-

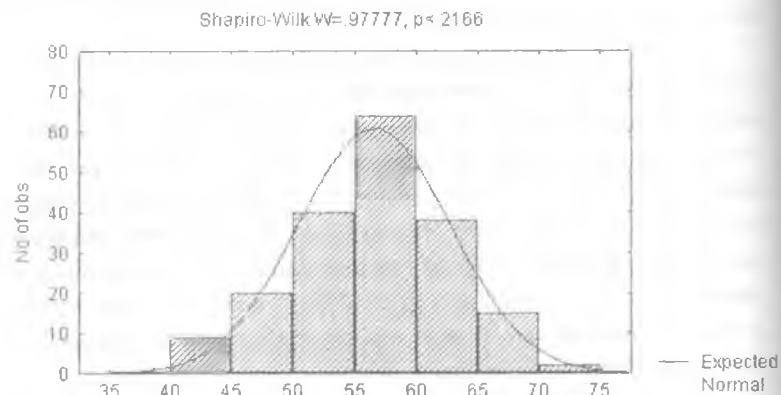


Рис.1. Гистограмма нормального распределения минимальных значений интенсивности фазы смычки [t]

* Здесь и далее приводятся примеры из работы "Реализация консонантных сочетаний на стыке лексических единиц (инструментально-фонетическое исследование на материале речи дикторов радио и телевидения Германии)" [5].

ленного на рис.1, и совпадает с наложенным на нее графиком для экспоненциального распределения. При этом точность совпадения оценивается с помощью критерия χ^2 или одновыборочного критерия Колмогорова-Смирнова [8].

Если в результате проведенного выше анализа установлено, что данная выборка подчиняется нормальным законам распределения, то, основываясь на этом можно отсеять **грубые ошибки измерений**, руководствуясь правилом *двух сигм* ($\bar{x} \pm 2\sigma$) и критерием Романовского, вычисляемым по формуле:

$$t_p = \frac{\bar{x}_i - \bar{x}}{\sigma} \quad (7)$$

Вычисленное значение сравнивается с табличным в соответствии с заданным объемом выборки и вероятностью погрешности [3, 13]. Если $t_p = t_{p'}$, то x_i отбрасывать нельзя. Если $t_p > t_{p'}$, то резко отличающееся значение x_i является грубой ошибкой и должно быть исключено из выборки.

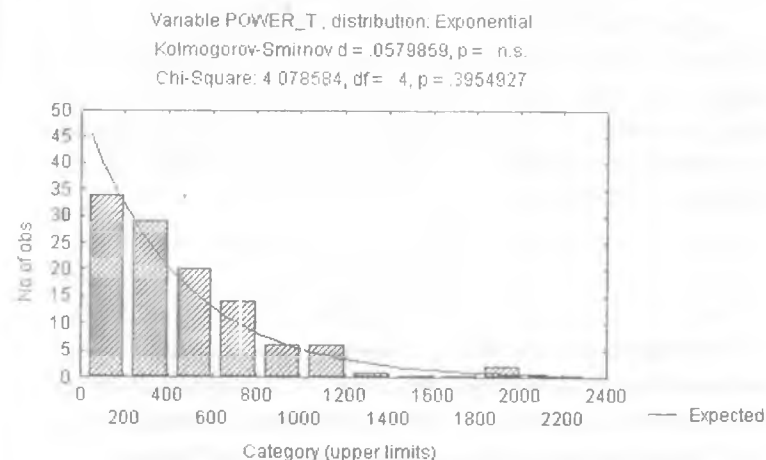


Рис. 2. Гистограмма экспоненциального распределения значений мощности смычного глухого [t]

При планировании инструментально-фонетического исследования необходимо исходить из того, что достоверная информация должна быть получена максимально экономным способом. Сущность закона больших чисел заключается в том, что по мере увеличения числа наблюдения, влияние случайных причин, определяющих величину признака у отдельных единиц лингвистической совокупности, взаимно погашается, и в сводных характеристиках совокупности проявляется действие основных причин, которые определяют закономерность фонетического явления.

Увеличение объема выборки повышает надежность результатов исследования, но это связано с трудностью обработки материала и значительной затратой времени. Поэтому **определение минимального объема выборки**, достаточного для описания лингвистического явления и обеспечивающего надежные, достоверные результаты, является центральным моментом в планировании фонетического исследования.

Для определения достаточного объема выборки можно использовать информацию о таких характеристиках, как среднее арифметическое, среднеквадратическое отклонение и коэффициент вариации, значения которых приближаются к истинным значениям всего процесса с определенной точностью ξ , определяемой по формулам:

$$\begin{aligned}\bar{X}_{ист} &= \bar{X}_{выб} \pm \xi \\ \sigma_{ист} &= \sigma_{выб} \pm \xi \\ \bar{V}_{ист} &= \bar{V}_{выб} \pm \xi\end{aligned}\quad (8)$$

Вероятность того, что данные выборочные статистические значения попадут в интервал значений $\bar{x} \pm \xi$, определяется уровнем статистической значимости p . Интервал значений, равный $\bar{x} \pm \xi$, называется *доверительным интервалом*. Точность ξ может быть задана в единицах измерения исследуемого признака, как в абсолютных, так и в относительных единицах.

Для определения объема выборки необходимо задать степень точности ξ и доверительную вероятность p . Длина доверительного интервала, т.е. заданная точность ξ определяется равенством:

$$\xi = Z_p \frac{\sigma}{\sqrt{n}} \quad (9)$$

где Z_p – аргумент функции интеграла Лапласа, определяемый по таблице [3].

Преобразуя формулу (9) относительно n устанавливаем, что объем выборки по заданной точности ξ и надежности p определяется следующей зависимостью:

$$n = Z_p^2 \frac{\sigma^2}{\xi^2} \quad (10)$$

В ходе исследования каждая из полученных в результате отбора выборок должна быть подвергнута проверке на достаточность.

Не менее важным в фонетических исследованиях является **определение количества дикторов**, достаточного для обеспечения репрезентативности выборки, содержащей изучаемое явление. Для этого используется понятие доверительных интервалов. При заданной доверительной вероятности $p = 95\%$ и имеющих значения выборочной средней $\bar{x}$ для всех полученных выборок можно получить доверительные интервалы по формуле (9).

Получив доверительные интервалы первых двух дикторов, сравниваем их с доверительными интервалами третьего диктора. Последние не укладываются в границы заданной точности, вычисленные по первым двум дикторам, из чего делаем вывод о недостаточности использования двух дикторов. Последовательно сравнивая доверительные интервалы следующих дикторов с доверительными интервалами суммированных предыдущих, находим количество дикторов, необходимое для данных выборок.

При сравнении показателей акустических характеристик звуков большое значение имеет вопрос о **существенности различий** между исследуемыми величинами. Как правило, среднее арифметическое значение показателя в одной выборке отличается от среднего арифметического значения в другой выборке.

В случаях, когда закон распределения этого показателя в обеих выборках близок к нормальному, а значения дисперсии существенно не отличаются друг от друга, ответ на поставленный вопрос дает *t*-критерий Стьюдента, который определяется по формуле:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(n_1 - 1) \cdot \sigma_1^2 + (n_2 - 1) \cdot \sigma_2^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \quad (11)$$

Полученное значение *t* затем сравнивается с табличным для заданной достоверности и числа степеней свободы [2]. Если полученное значение больше табличного, то различия между двумя средними значениями являются статистически значимыми. Если полученное значение меньше табличного, то различия не являются существенными и ими можно пренебречь.

Современные программные пакеты статистического анализа позволяют не только вычислять значение *t* для зависимых и независимых выборок, но и рассчитывать степень достоверности различий между двумя средними для данного количества степеней свободы. Если полученное в результате $p \leq 0,05$, то различия можно считать существенными. Полученные значения *p* в пределах 0,005-0,001 свидетельствуют о высокой значимости различий между двумя переменными. Следовательно, чем выше *p*-уровень, тем меньше мы можем доверять наблюдаемым связям между соответствующими переменными выборки.

Как указывалось выше, важной предпосылкой истинности *t*-теста является однородность дисперсии, которая проверяется по тесту Левена, основанному на абсолютном отклонении значений от среднего арифметического. Если результаты теста являются статистически значимыми, т.е. $p \leq 0,05$, то предпосылка об однородности дисперсии должна быть отклонена [8].

В этом случае, как и в случае, когда хотя бы одна из выборок не соответствует нормальному закону распределения, следует применять один из непараметрических эквивалентов *t*-теста Стьюдента для независимых выборок, а именно, *U*-тест Манна-Уитни, действие которого основано на ранговых суммах, а не на среднеарифметических значениях. Это самый чувствительный из непараметрических эквивалентов *t*-теста Стьюдента для независимых выборок. Толкование этого теста аналогично толкованию *t*-теста Стьюдента, а именно, если $p \leq 0,05$, то различия можно считать существенными, в противном случае ими можно пренебречь [8].

Для установления общих законов, лежащих в основе определенного фонетического явления, следует проанализировать не только каждый из компонентов какого-либо явления, но и представить **детальную количественную и качественную характеристику связей**, существующих между любыми двумя признаками процесса с последующим описанием результатов воздействия данных связей на изучаемые явления.

В случае если обе выборки подчиняются законам нормального распределения, то взаимное влияние двух показателей можно описать с помощью коэффициента *линейной корреляции Пирсона* [4], который определяется по формуле:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (12)$$

Получение числового коэффициента не всегда является достаточным для правильного определения связи между параметрами, поскольку он не отражает наличия в выборке так называемых *выбросов*, т.е. резко отклоняющихся значений. Наличие хотя бы одного из них может в значительной мере исказить реальную картину связи между величинами [8]. Так, например, при рассмотрении графика корреляционной зависимости интенсивности фазы смычки и ее длительности (рис.3) из него были ис-

ключены величины, резко отличающиеся от остальных, что позволило установить точный уровень корреляции между временным и энергетическим параметрами.

Если одна из двух сравниваемых выборок не подчиняется законам нормального распределения, необходимо использовать один из непараметрических эквивалентов r -критерия Пирсона.

Таким эквивалентом является коэффициент *ранговой корреляции Спирмана*, определяемый по формуле:

$$R = 1 - \frac{6 \cdot \sum (x_i - y_i)^2}{n \cdot (n^2 - 1)} \quad (13)$$

На рис. 4 представлен график зависимости параметра мощности глухого предстыкового плюзива [t] от уровня интенсивности фазы смычки, полученный в результате проведенного рангового корреляционного анализа Спирмана. Применение этого непараметрического теста вызвано тем, что один из сравниваемых параметров подчиняется экспоненциальному закону распределения.

В целом, схема статистического анализ данных инструментального анализа может быть представлена следующим образом (рис.5).

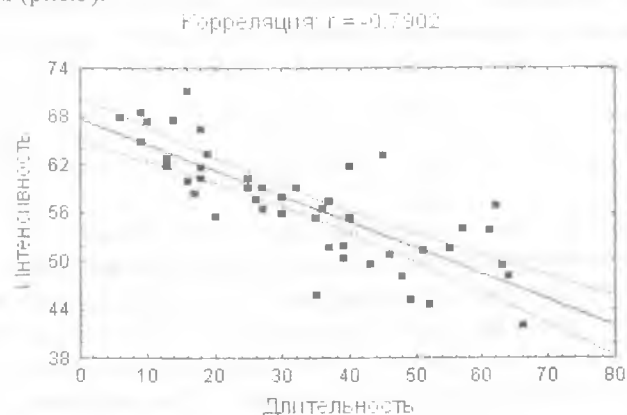


Рис. 3. График корреляционной зависимости интенсивности фазы смычки [t] и ее длительности по r -критерию Пирсона

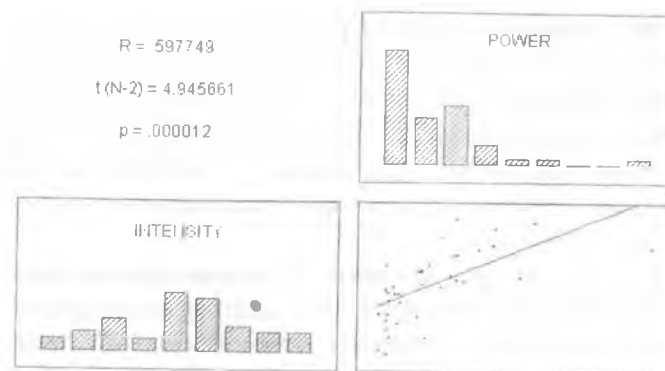


Рис. 4. График корреляционной зависимости мощности глухого предстыкового плюзива [t] от уровня интенсивности фазы смычки по R -критерию Спирмана

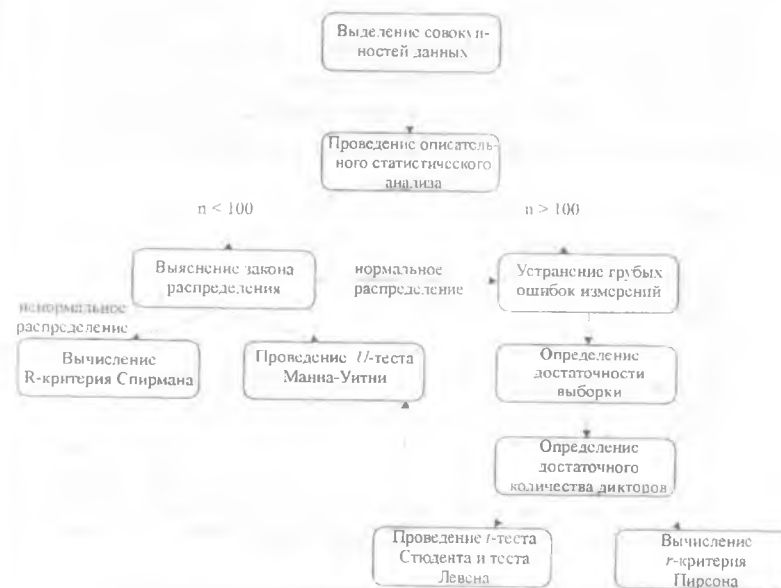


Рис.5. Схема статистического анализа данных инструментального исследования

При обработке данных **перцептивного** анализа можно использовать метод *перекрестной табуляции*, в основе которого лежит сравнение двух или более таблиц распределения номинативных и ординальных величин [8]. Изучение полученных частотных данных дает представление о частотном распределении таких изучаемых фонетических характеристик как аспирация, паузация и др.

Наиболее распространенным методом проверки взаимоотношений между номинативными и ординальными переменными является критерий χ^2 Пирсона, с помощью которого оценивается существенность различий между нулевой гипотезой и гипотезой, выдвинутой на основании полученных данных перцептивного анализа. Если результаты этого теста являются статистически значимыми, то нулевую гипотезу об отсутствии связи следует отклонить. При этом уровень статической значимости определяет величину связи между параметрами. Результаты перекрестно-табуляционного анализа, например взаимоотношений между наличием/отсутствием паузы в звукосочетании и характером аспирированности пловиза, представлены в виде графика на рис.6.

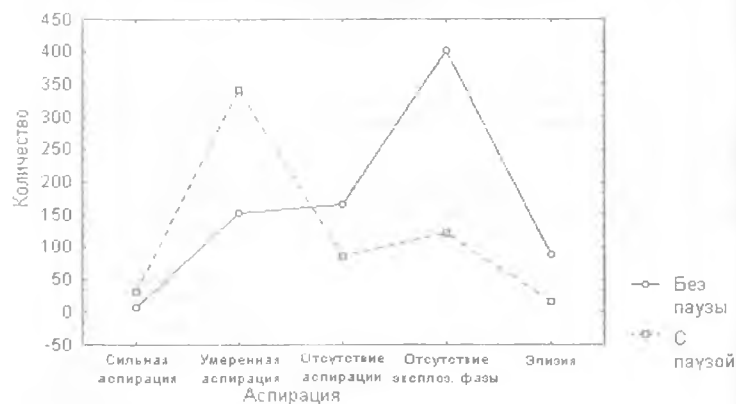


Рис.6. График результатов перекрестно-табуляционного анализа аспирации предстыковых пловизов в зависимости от наличия/отсутствия паузы

Таким образом, использование методов статистической обработки данных позволяет учесть все особенности исследуемого речевого материала и адекватно отразить существующие связи между звуковыми явлениями независимо от характера его распределения в выборке.

Литература

1. Бектаев К.Б., Белоцерковская Л.И., Пиотровский Р.Г. Норма – ситуация – текст и лингвостатистические приемы их исследования. // Языковая норма и статистика. – М.: Наука, 1977. – 301 с.
2. Бровченко Т.А., Варбанец П.Д., Таранец В.Г. Метод статистического анализа в фонетических исследованиях. – Одесса: Изд-во ОГУ, 1976. – 101 с.
3. Бровченко Т.А., Волошин В.Г. Методические указания по математической обработке и анализу результатов фонетического эксперимента. – Одесса: ОГУ, 1986. – 49 с.
4. Носенко И.А. Начала статистики для лингвистов. – М.: Высшая школа, 1981. – 157 с.
5. Петлюченко Н.В. Реализация консонантных сочетаний на стыке лексических единиц (инструментально фонетическое исследование на материале речи дикторов радио и телевидения Германии): Дис. ... канд. филол. наук: 10.02.04. – Одесса, 1999. – 230 с.
6. Чернова Н.И. Краткий конспект лекций по математической статистике. – Новосибирск, 1998 – <http://www.nsu.ru/mmftvims/chernova/ms>
7. Muller Ch. Einführung in die Sprachstatistik. – Berlin, 1972.
8. StatSoft, Inc. Electronic Statistics Textbook. – Tulsa, OK: StatSoft, 1997.