文章编号:1004-9037(2014)02-0227-05

基于 GMM 和 ANN 混合模型的语音转换方法

姚绍芹 张玲华

(南京邮电大学通信与信息工程学院,南京,210003)

摘要:为了克服利用高斯混合模型(Gaussian mixture model, GMM)进行语音转换的过程中出现的过平滑现象,考虑到 GMM 模型参数的均值能够表征转换特征的频谱包络形状,提出一种基于 GMM 与人工神经网络(Artificial neural network, ANN)混合模型的语音转换。该方法利用 ANN 对 GMM 模型参数的均值进行转换;为了获取连续的转换频谱,采用静态和动态频谱特征相结合来逼近转换频谱序列;鉴于基频对语音转换的重要性,在频谱转换的基础上,对基频也进行了分析和转换。最后,通过主观和客观实验对提出的混合模型的语音转换方法的性能进行测试。实验结果表明,与传统的基于 GMM 模型的语音转换方法相比,本文提出的方法能够获得更好的转换语音。

关键词:频谱转换;高斯混合模型;径向基函数;神经网络

中图分类号:TN912.3 **文献标志码**:A

Voice Conversion Based on Mixed GMM-ANN Model

Yao Shaoqin, *Zhang Linghua*

(College of Telecommunication & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China)

Abstract: As the mean vector of Gaussian mixture model (GMM) parameters can represent the basic shapes of converted feature vectors, based on a mixed model comprised of GMM and artificial neural network (ANN), a novel spectral conversion method is proposed. The method alleviates the over-smoothing problem by using ANN to transform the mean vector of GMM parameters. Static and dynamic spectral features are used for approaching the converted spectrum sequence in order to gain the continuous converted spectral. Moreover, as pitch is very important to voice conversion, it is also analyzed and transformed on the basis of spectral conversion. The performance of the proposed method is evaluated using subjective and objective tests, and the results show that the proposed method can obtain a better speech quality than the earlier voice conversion system based on conventional GMM method.

Key words: spectral conversion; Gaussian mixture model; radial basis function; neural network

引 言

语音转换[1]试图对源说话人的语音进行转换,使其听起来像是目标说话人说的一样。语音转换应用于多个领域,比如电影配音、文语合成、医疗康复等。

设计语音转换系统,最基本的问题在于声学特征的选择。众所周知,早期的语音转换系统主要集中在频谱包络的转换上,这是因为频谱包络在提取源说话人语音特征方面发挥至关重要的作用。然而除此之外,一些韵律特征,如基音对获取高质量的合成语音起着至关重要的作用。

事实上,矢量量化(Vector quantization, VQ)[2]、高斯混合模型(Gaussian mixture model, GMM)[3-6]、人工神经网络(Artificial neural network, ANN)[5,7,8]等多种方法已经被用于源语音特征矢量到目标语音特征矢量的映射以获取转换

基金项目:江苏省高校自然科学研究(13KJA510003)重大资助项目;江苏高校优势学科建设工程(PAPD)资助项目;江苏省普通高校研究生科研创新计划(CXLX12_0478,CXZZ13_0488)资助项目。

收稿日期:2014-01-02;**修订日期**:2014-01-09

函数。其中,GMM 凭借良好的性能得以广泛应用。尽管如此,GMM 中经常出现的过平滑现象依旧极大地降低了转换语音的质量。鉴于此,文献[9]通过 MAP(Maximum a posteriori)自适应,文献[10]采用转换频谱的全局变量特征,均试图探索解决高斯混合模型中的过平滑问题。

本文认为 GMM 模型中均值矢量是生成转换语音的基本包络形状,因此,通过改进均值矢量来缓解过平滑现象带来的影响。由于源说话人和目标说话人在声道上的变化是非线性的,且基于 ANN 模型的语音转换模型与基于 GMM 模型的转换语音在效果上不分伯仲[5],基于 ANN 和 GMM 的混合模型应运而生,即采用 ANN 模型对 GMM 模型中的均值矢量进行映射。ANN 模型包含很多种类型,本文拟采用径向基函数(Radial basis function, RBF)神经网络,这是因为它拥有快速的训练过程,并且能够以比较简单的网络架构实现更精确的逼近。

由于语音的韵律特征,尤其是基频 F_0,包含了大量的说话人的个性特征,同时考虑到频谱特征与基频的相关性以及特征之间的非线性[11],本文将采用基于 RBF 神经网络的联合频谱特征参数的基频转换[12]。

1 特征参数的提取

频谱包络特征可以由多种特征矢量来表示。本文采用 16 阶线谱频率(Line spectrum frequency, LSF)[13],这是因为 16 阶 LSF 能更好地表征声道与共振峰模型,并具有良好的插入特性。此外,考虑到 LSF 具有较强的帧间相关性,因此,为了获得连续的转换频谱,16 阶 LSF 的动态特征 Δ 也被用来与 16 阶静态 LSF 一起形成 32 阶特征矢量表示频谱特征。其中,动态特征 Δ 指的是相邻帧间的差值。动态时间规整(Dynamic time warping, DTW)用来对齐源语音与目标语音的特征矢量。此外,由于语音的韵律特征,尤其是 F_0,包含了大量的说话人的个性特征,因此,本文也对基频进行了转换。而在分析与合成阶段,STRAIGHT 模型[14]用来提取语谱参数和 F_0,对语谱参数进行快速傅里叶逆变换得到自相关系数,对自相关系数进行 Levinson-Durbin 算法得到自回归参数,即线性预测系数(Linear prediction coefficient, LPC)系数,最后,由 LPC 系数转换成 LSF 系数。

2 基于 GMM 模型的频谱转换算法

假设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$ 和 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\}$ 分别表示 N 个时间对齐的源说话人和目标说话人的频谱特征矢量,其中,N 表示语音的帧数,矢量 $\boldsymbol{x}_t$(或是 $\boldsymbol{y}_t$)是 p 维特征矢量。语音转换可以理解成将源特征矢量 $\boldsymbol{x}_t$ 转换成目标特征矢量 $\boldsymbol{y}_t$ 的过程。通过最小化转换特征矢量 $\hat{\boldsymbol{y}}_t=F(\boldsymbol{x}_t)$ 与目标特征矢量 $\boldsymbol{x}_t$ 在所有帧间的平方误差总和,进一步得到映射函数 F。

在早期的研究中,主要有两类基于 GMM 模型的语音转换方法,即源 GMM 模型方法[3]和联合密度模型方法[4]。两种方法性能上相差无几,本文采用后者作为基本的频谱转换方法。

在联合密度模型中,联合特征矢量 $\boldsymbol{Z}$ 表示源与目标特征矢量的集合 $\boldsymbol{Z}=(\boldsymbol{X}^{\mathrm{T}},\boldsymbol{Y}^{\mathrm{T}})^{\mathrm{T}}$,其中 T 代表矢量的转置,然后利用联合特征矢量 $\boldsymbol{Z}$ 对 GMM 模型进行训练。源与目标特征矢量的联合概率密度函数表示如下

$$p(\boldsymbol{z})=\sum_{i=1}^{M}\alpha_i N(\boldsymbol{z};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\quad\sum_{i=1}^{M}\alpha_i=1,\ \alpha_i\geqslant 0 \tag{1}$$

式中:$N(\boldsymbol{z};\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$表示均值为 $\boldsymbol{\mu}_i$、协方差矩阵为 $\boldsymbol{\Sigma}_i$ 的正态分布;α_i 表示第 i 个高斯分量的先验概率;M 表示高斯分量的总数目。通过最大期望(Expectation maximization, EM)迭代算法估算 GMM 参数($\boldsymbol{\alpha}_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$)。均值 $\boldsymbol{\mu}_i$ 和协方差 $\boldsymbol{\Sigma}_i$ 可以表示为

$$\boldsymbol{\mu}_i=\begin{bmatrix}\boldsymbol{\mu}_i^x\\ \boldsymbol{\mu}_i^y\end{bmatrix},\ \boldsymbol{\Sigma}_i=\begin{bmatrix}\boldsymbol{\Sigma}_i^{xx} & \boldsymbol{\Sigma}_i^{xy}\\ \boldsymbol{\Sigma}_i^{yx} & \boldsymbol{\Sigma}_i^{yy}\end{bmatrix} \tag{2}$$

高斯混合模型的转换函数为

$$F(\boldsymbol{x})=E[\boldsymbol{y}\mid\boldsymbol{x}]=\sum_{i=1}^{M}h_i(\boldsymbol{x})[\boldsymbol{\mu}_i^y+\boldsymbol{\Sigma}_i^{yx}(\boldsymbol{\Sigma}_i^{xx})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i^x)] \tag{3}$$

式中:$h_i(\boldsymbol{x})$表示给定的输入矢量 $\boldsymbol{x}$ 属于第 i 个高斯分量的后验概率,如式(4)所示。

$$h_i(\boldsymbol{x})=\frac{\alpha_i N(\boldsymbol{x};\boldsymbol{\mu}_i^x,\boldsymbol{\Sigma}_i^{xx})}{\sum_{m'=1}^{M}\alpha_{m'}N(\boldsymbol{x};\boldsymbol{\mu}_{m'}^x,\boldsymbol{\Sigma}_{m'}^{xx})} \tag{4}$$

3 基于混合模型的频谱转换算法

尽管基于 GMM 模型的频谱转换得到了广泛使用,但是它依旧受制于过平滑现象,并且无法获得声道特性间的非线性关系。从式(3)中可以得知,映射函数包含两部分组成,其中均值矢量 $\boldsymbol{\mu}_i^y$ 代表转换特征的基本频谱包络形状。为了解决过平

滑问题，拟考虑使用 RBF 神经网络用于对均值矢量 $\boldsymbol{\mu}_i^y$ 进行转换。

3.1 RBF 神经网络

RBF 神经网络[15] 是由 Broomhead 和 Lowe 提出的前馈网络。RBF 神经网络包含 3 层：即输入层、隐层和输出层。输入层不作转换，仅仅将输入特征矢量分派到隐层。隐蔽层采用径向基函数，将输入特征矢量转换到隐层空间。输出层主要实现对隐蔽层的输出加权求和。

RBF 神经网络通过将源说话人的声学特征转换到目标说话人的声学特征来获取转换函数。如果 $\hat{\boldsymbol{y}}_j$ 代表矢量 $\boldsymbol{x}$ 通过 RBF 神经网络映射后的输出，那么 $\hat{\boldsymbol{y}}_j$ 为

$$\hat{\boldsymbol{y}}_j = \sum_{i=1}^{N} \boldsymbol{w}_{ij} \phi_i(\boldsymbol{x}) \quad 1 \leqslant j \leqslant m \tag{5}$$

式中：N 表示径向基函数的数目；$\boldsymbol{w}_{ij}$ 表示输出层的权值；m 表示输出特征矢量的维数；$\phi_i(\boldsymbol{x})$ 表示径向基函数（高斯函数），如下所示

$$\phi_i(\boldsymbol{x}) = \exp\left(-\frac{\| \boldsymbol{x} - \boldsymbol{c}_i \|^2}{2\boldsymbol{\sigma}_i^2}\right) \quad 1 \leqslant i \leqslant N \tag{6}$$

式中 $\boldsymbol{c}_i$ 和 $\boldsymbol{\sigma}_i^2$ 分别表示隐层 RBF 的中心和宽度。

3.2 基于 GMM 与 ANN 的混合模型的频谱转换算法

本文提出的混合语音转换方法包含两个阶段，即训练阶段和转换阶段：

（1）训练阶段

第一步：针对如前所述的源与目标联合矢量集合 $\boldsymbol{Z}$，采用 EM 算法确定 GMM 参数序列（$\boldsymbol{\alpha}_i$，$\boldsymbol{\mu}_i$，$\boldsymbol{\Sigma}_i$）。根据式（3），进一步确定 GMM 映射函数。

第二步：生成用于 RBF 神经网络训练的数据集合

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{\mu}_i^x \tag{7}$$

$$\boldsymbol{y}_{\text{new}} = \boldsymbol{\mu}_i^y \tag{8}$$

第三步：依据输入 $\boldsymbol{x}_{\text{new}}$ 和输出 $\boldsymbol{y}_{\text{new}}$，构造 RBF 神经网络映射函数 F_{rbf}。

（2）转换阶段

第一步：针对测试矢量 $\boldsymbol{X}'$，采用 EM 算法估算 GMM 参数序列 $\boldsymbol{\mu}'_i$。

第二步：依据 RBF 神经网络映射函数 F_{rbf}，得到新的均值矢量 $\boldsymbol{\mu}'_{\text{new},i}$

$$\boldsymbol{\mu}'_{\text{new},i} = F_{\text{rbf}}(\boldsymbol{\mu}'_i) \tag{9}$$

第三步：采用新的均值矢量 $\boldsymbol{\mu}'_{\text{new},i}$ 代替 $\boldsymbol{\mu}_i^x$，得到新的 GMM 映射函数。

第四步：依据新的 GMM 映射函数，得到转换后的频谱特征矢量。

3.3 基频转换

由于语音的韵律特征，尤其是 F_0，包含了大量的说话人的个性特征，同时考虑到频谱特征与基频的相关性以及特征之间的非线性，本文将采用基于 RBF 神经网络的联合频谱特征参数的基频转换。在训练阶段，首先对用于训练的目标语音进行分帧处理及清浊音判断，利用 STRAIGHT 模型依照第一部分的特征参数提取的方法对浊音帧提取频谱特征和 F_0，鉴于女声的基频范围在 60～450 Hz，男声的基频范围在 60～200 Hz，RBF 网络的输出要求在 0～1 之间，因此，必须对提取的 F_0 除以 500 进行缩放，然后将频谱特征与缩放后的 F_0 分别作为 RBF 神经网络的输入和输出，从而得到转换函数。在转换阶段，首先提取待转换语音的频谱，然后采用第三部分获取的新的 GMM 映射函数对其进行转换得到转换的频谱，然后根据训练阶段获得的转换函数对转换的频谱进行映射得到缩放后的转换基频，再将其乘以 500，即可获得最终的转换基频。

4 实验与讨论

本实验通过主观和客观测试进一步检验所提方法的性能。鉴于听力测试对于频谱转换算法的性能评估至关重要，拟采用平均意见分（Mean opinion score，MOS）和 ABX 测试完成频谱转换系统的主观评价，而客观评价主要以频谱失真为评价依据。GMM 采用 20 个高斯分量。实验在一个平行语料库里完成。语料库包含 141 个汉字和 6 个短句子，它们分别来自于两个男声和两个女声。所有音频的采样频率均为 16 kHz，以 16 bit 量化。随机选取 100 个汉字作为训练数据，其他全部用于测试。其中，只有浊音用于训练与转换，清音保持不变。并且，实验主要以异性转换为基础，包含男声向女声转换和女声向男声转换。

4.1 主观评价

实验主要采用两种不同的主观方法来验证所提算法的实际性能，即 ABX 测试以及 MOS 测试。

ABX 测试用于评价目标语音与转换语音的近似度。假设 A 和 B 分别代表源说话人语音和目标说话人语音，X 代表采用了上述 2 种方法转换而来的语音。实验要求 10 位经验丰富的听众从转换的语音中选择 A 或 B 哪一个听起来最接近 X，共 40 个汉字需要听众们一一评价。表 1 显示 4 种转换

方法的 ABX 测试结果。

表 1 ABX 测试结果

Table 1 Result of ABX test

方法	正确率/%			
	M-M	M-F	F-F	F-M
GMM	47.26	53.65	48.3	50.8
GMM+RBF	58.39	66.54	60.4	62.9

注:M-M 表示男声-男声;M-F 表示男声-女声;F-F 表示女声-女声;F-M 表示女声-男声

从表 1 可以看出,本文提出的混合算法效果明显优于传统的基于 GMM 模型的语音转换方法,同时,异性间的转换也比同性间转换更接近目标说话人的语音,尤其是男声到女声的转换更为突出,从之前的 53.65%提高到 66.54%,提升了 12.89%。

MOS 测试是另一种主观测试方法,它同样要求 10 位经验丰富的听众采用 5 分制依次为转换后的语音的质量进行打分(1:非常差;2:较差;3:一般;4:较好;5:非常好)。实验结果如表 2 所示。

表 2 MOS 测试结果

Table 2 Result of MOS test

方法	M-M	M-F	F-F	F-M
GMM	2.65	2.53	2.74	2.44
GMM+RBF	2.75	2.67	2.85	2.51

根据表 2 的实验结果,可以得出的结论是:本文提出的语音转换方法比传统的基于 GMM 的语音转换方法性能更佳,同时由于同性间的个性差异较小,所以其转换性能要优于异性间的转换。对于异性间的转换而言,男声到女声的转换效果也要好于女声到男声的转换。

4.2 客观评价

频谱失真(Spectral distortion, SD)是一种常见的频谱转换客观评价方法,如式(10)所示。

$$\mathrm{SD}=\frac{1}{N}\sum_{i=1}^{N}\frac{\|\boldsymbol{y}_i-F(\boldsymbol{x}_i)\|}{\|\boldsymbol{y}_i-\boldsymbol{x}_i\|} \tag{10}$$

式中:$\boldsymbol{x}_i$,$\boldsymbol{y}_i$ 和 $F(\boldsymbol{x}_i)$分别表示源说话人的特征矢量、目标说话人的特征矢量和转换的特征矢量;N 代表语音帧数。图 1~2 分别显示了女声到男声转换和男声到女声转换的频谱失真情况。

从图 1~2 可以看出,本文提出的方法的谱失真率明显小于传统的基于 GMM 模型的语音转换的谱失真率,即转换的性能更优,同时,该方法在男声到女声的转换中效果更佳。

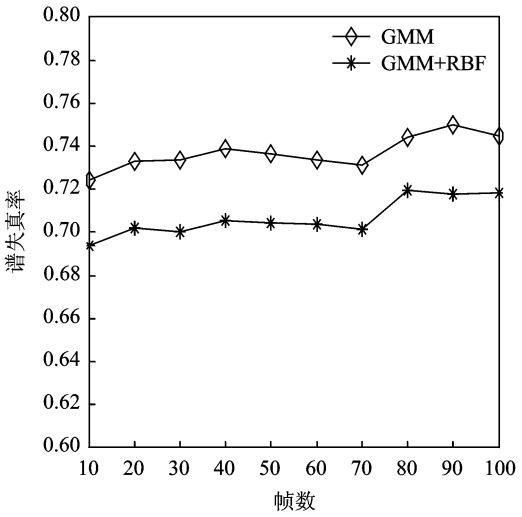

图 1 女声到男声的语音频谱失真图

Fig. 1 Spectral distortion (F-M)

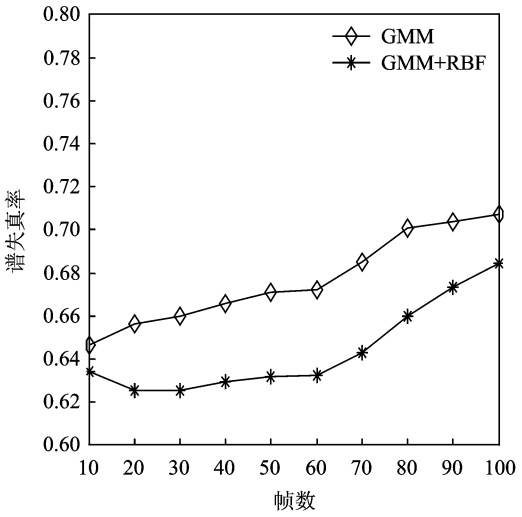

图 2 男声到女声的语音频谱失真图

Fig. 2 Spectral distortion (M-F)

5 结束语

考虑到 GMM 模型参数的均值能够表征转换特征的频谱包络形状,本文提出一种基于 GMM 与 ANN 的混合模型的语音转换方法来克服利用 GMM 进行语音转换的过程中出现的过平滑现象,主要做法就是利用 RBF 神经网络对 GMM 模型参数的均值进行转换以获得新的 GMM 模型的转换函数。同时,考虑到 LSF 具有较强的帧间相关性,为了获取连续的转换频谱包络,采用了静态和动态频谱特征相结合来逼近转换频谱序列。此外,由于基频对于高质量的语音转换至关重要,同时考虑到频谱特征与基频之间的相关性,因此,在频谱转换

的基础上，采用了 ANN 模型对基频也进行了转换。最后，通过主观和客观实验对提出的转换方法的性能进行测试，实验结果表明与传统的基于 GMM 的方法相比，本文提出的方法能够获得更好的转换语音。

参考文献：

[1] 孙健，张雄伟，曹铁勇，等. 基于卷积非负矩阵分解的语音转换方法[J]. 数据采集与处理，2013，28(1)：141-148.
Sun Jian，Zhang Xiongwei，Cao Tieyong，et al. Voice conversion based on convolutive nonnegative matrix factorization[J]. Journal of Data Acquisition and Processing，2013，28(1)：141-148.

[2] Abe M，Nakamura S，Shikano K，et al. Voice conversion through vector quantization[C]// IEEE International Conference on Acoustics，Speech and Signal Processing. New York，USA：IEEE，1988：655-658.

[3] Stylianou Y，Cappe O，Moulines E. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing，1998，6(2)：131-142.

[4] Kain A，Macon M W. Spectral voice conversion for text-to-speech synthesis [C]//IEEE International Conference on Acoustics，Speech and Signal Processing. Seattler，WA，USA：IEEE，1998：285-288.

[5] Laskar R H，Chakrabarty D，Talukdar F A，et al. Comparing ANN and GMM in a voice conversion framework[J]. Applied Soft Computing，2012，12(11)：3332-3342.

[6] 岳振军，邹翔，王浩. 基于隐马尔可夫模型和高斯混合模型结合的声音转换方法[J]. 数据采集与处理，2009，24(3)：285-289.
Yue Zhenjun，Zou Xiang，Wang Hao. Voice conversion with the combination of HMM and GMM[J]. Journal of Data Acquisition and Processing，2009，24(3)：285-289.

[7] Desai S，Black A W，Yegnanarayana B，et al. Spectral mapping using artificial neural networks for voice conversion[J]. IEEE Transactions on Audio，Speech and Language Processing，2010，18(5)：954-964.

[8] Rao K S. Voice conversion by mapping the speaker-specific features using pitch synchronous approach [J]. Computer Speech and Language，2010，24(3)：474-494.

[9] Chen Yining，Chu Min，Chang Eric，et al. Voice conversion with smoothed GMM and MAP adaptation [C]// 8th European Conference on Speech Communication and Technology. Geneva，Switzerland：ISCA Archive，2003：2413-2416.

[10] Toda T. Black A W，Tokuda K. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory[J]. IEEE Transactions on Audio，Speech and Language Processing，2007，15(8)：2222-2235.

[11] Shao Xu，Milner Ben. Pitch prediction from MFCC vectors for speech reconstruction[C]// IEEE International Conference on Acoustics，Speech and Signal Processing. Montreal，Que，Canada：IEEE，2004：97-100.

[12] 解伟超. 语音转换中声道谱参数和基频变换算法的研究[D]. 南京：南京邮电大学，2013.
Xie Weichao. The research on vocal tract spectrum and pitch frequency transformation in voice conversion[D]. Nanjing：Nanjing University of Posts and Telecommunications，2013.

[13] Turk O，Arslan L M. Robust processing techniques for voice conversion[J]. Computer，Speech and Language，2006，20(4)：441-467.

[14] Kawahara H，Masuda-Katsuse I，de Cheveigné A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction：Possible role of a repetitive structure in sounds[J]. Speech Communication，1999，27(3/4)：187-207.

[15] Watanabe T，Murakami T，Namba M，et al. Transformation of spectral envelope for voice conversion based on radial basis function network[C]//7th International Conference on Spoken Language Processing. Denver，Calorado，USA：ISCA Archive，2002：285-288.

作者简介：姚绍芹(1988-)，女，博士研究生，研究方向：语音转换，E-mail：yaoshaoqin000@163.com；张玲华(1964-)，女，教授，研究方向：语音信号处理、智能信号处理、语音通信、无线通信中的信号处理。