SottoVoce：基于超声成像的深度神经网络无声语音交互

Naoki Kimura 东京大学 7-3-1, Hongo, Bunkyo-ku, 东京 日本
ki
mura-naoki@g.ecc.u-tokyo.ac.jp

Michinari Kono 东京大学 7-3-1, Hongo, Bunkyo-ku, 东京 日本
mchkono@acm.org

Jun Rekimoto*东京大学 7-3-1,Hongo, Bunkyo-ku, 东京 日本 索尼计算机科学实验室 3-14-13 Higashigotanda, 品川区, 东京,日本 rekimoto@acm.org

arXiv:2303.01758v1 [cs.HC] 3 Mar 2023

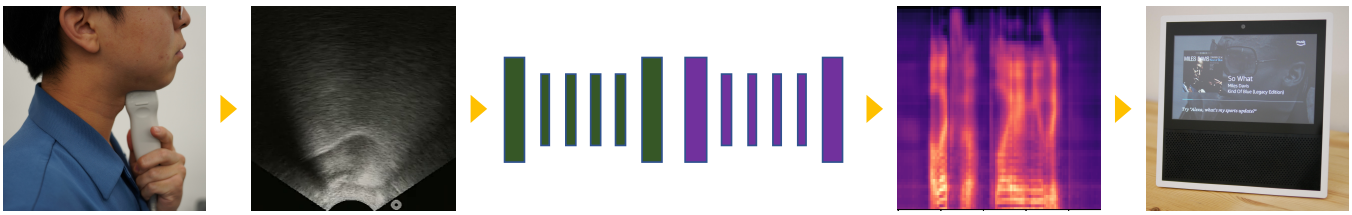

图 1：SottoVoce 无声语音系统：将超声回波探头贴在下颌下方，在用户说话时读取其内部状况，而用户实际上并未发出声音。通过使用深度卷积神经网络识别超声图像，系统可重新合成用户的语音，并将其用于控制智能音箱等现有语音交互系统。

摘要

通过语音操作的数字设备日益普及。然而，语音界面的应用仍受到限制。例如，在公共场所说话会打扰周围的人，而且不应说出机密信息。环境噪声可能会降低语音识别的准确性。为解决这些局限，本文提出了一种检测用户无声发音的系统。该系统通过安装在下颌下方的超声成像传感器获取内部信息，无需用户发出声音即可识别话语内容。我们提出的深度神经网络模型用于从超声图像序列中提取声学特征。我们确认，该系统生成的音频信号可以控制现有的智能音箱。我们还发现，用户可以调整口腔运动，通过练习提高语音识别的准确性。

CCS 概念

• **计算方法** → **神经网络**；• **以人为中心的计算** → 界面设计原型；移动设备；基于声音的输入／输出；• **社会与专业主题** → 残障人士。

关键词

无声语音，超声成像，深度神经网络，人机智能融合

ACM 参考格式：
Naoki Kimura、Michinari Kono 和 Jun Rekimoto。2019。SottoVoce：基于超声成像的无声语音交互：使用深度神经网络。载于《*CHI* 人机交互计算系统会议论文集》（*CHI 2019*），*2019*年*5*月*4*日至*9*日，英国苏格兰格拉斯哥。ACM，纽约，NY，美国，10页。https://doi.org/10.1145/3290605.3300376

*通讯作者

允许出于个人或课堂用途免费制作本作品全部或部分内容的数字或纸质副本，前提是这些副本不得用于营利或商业用途，且须保留本声明和首页上的完整引文。对于本作品中归属 ACM 以外其他方所有的组成部分，必须尊重其版权。允许在注明出处的情况下摘录。以其他方式复制、再版、发布到服务器或重新分发至列表，均须事先获得明确许可和／或支付费用。请向 permissions@acm.org 申请许可。*CHI 2019*，*2019*年*5*月*4*日至*9*日，英国苏格兰格拉斯哥© 2019 计算机协会。ACM ISBN 978-1-4503-5970-2/19/05…$15.00 https://doi.org/10.1145/3290605.3300376

1 引言

可通过语音控制的智能设备（即语音交互设备）已应用于许多场景。智能手机、智能音箱、汽车导航系统和各种家用电器都可以通过语音控制，展现出诸多交互潜力 [47]。语音交互无需用户注视屏幕，因此可以在卧室等黑暗环境中使用。得益于语音识别技术的近期进展以及语音合成的自然度不断提升，语音交互正成为人机交互中不可或缺的一环。用户在执行其他任务时，例如开车、做饭、居家办公或使用传统个人电脑，也可以使用语音交互设备。举例来说，当用户

专注于电脑屏幕，并使用键盘、鼠标等交互设备时，仍然可以通过语音交互操作其他设备。

然而，语音界面仍有两项挑战需要克服。首先，在公共场所使用该界面会受到限制。除了会打扰周围的人，在公共场所说出个人信息或机密信息，从信息安全角度来看也存在风险。其次，在嘈杂环境中无法使用该界面，因为语音识别的准确率可能会下降。尝试通过语音界面与可穿戴设备或移动计算机交互时，这些问题尤为突出。

为克服这些挑战，人们开展了关于“无声语音识别”的研究 [8]。例如，采用唇读这一方法时，摄像机会拍摄说话者的口腔图像或整张脸，再根据这些图像估计话语内容 [52]。如果用户只需用口腔做出话语的发音动作而不实际发声，就能在公共场所进行这类语音交互。然而，摄像头方法需要在脸部前方安装摄像机，其外形尺寸使其不适用于可穿戴设备或移动应用。其他研究则尝试通过佩戴在用户皮肤或喉部的麦克风或加速度计，识别“非听觉喃语”（NAM） [22, 42] 中的话语。在这种情况下，用户以清晰发音的呼吸声说话，但不产生声带振动（即耳语）。然而，为了让系统准确识别，用户的耳语往往会被附近的人注意到。此外，一些研究尝试通过肌电图（EMG） [33, 49] 估计口腔附近肌肉的运动，从而推断语音。然而，利用肌电图估计自由话语仍然很困难；相反，这属于利用口腔运动进行的一种手势识别。因此，可检测的命令数量有限，用户必须学习新的手势技能，而不能使用已有的说话技能。

我们不采用上述方法，而是专注于利用超声成像 [11]。超声成像技术通过测量向体内辐射的超声波的反射时间，识别身体内部的状况。这项技术广泛用于医疗领域，以了解内脏器官的状况。近年来，出现了可直接连接智能手机的小型轻量系统（例如通用电气公司的 Vscan Extend）。如果能够将小型超声成像探头佩戴在颈部周围，以感知口腔内的情况并将其转换为声学信息，那么它将成为一种实用设备，使人们无需实际出声即可与支持语音的设备通信；也就是说，无声语音交互将成为可能。

基于超声成像的无声语音交互相较于其他方法展现出两项潜在优势。首先，超声成像头可以小型化，并可制成项圈等外形不显眼的设备。这将是设计可穿戴无声语音系统的一项重要特性。其次，通过识别口腔内的情况，可以测量舌头的运动，而这无法从外部观察到。因此，有望更准确地还原声音。

已有研究利用超声成像进行无声语音研究，但其中许多方法需要结合唇部或面部图像，因此必须在用户 [26] 前方放置摄像机。将其用作可穿戴界面设备时，这种配置存在局限。近期研究人员尝试将深度神经网络与超声成像结合用于无声语音 [7, 51]；然而，这些方法并非基于卷积神经网络，也未使用从未发声时的口腔运动中采集的数据进行验证。它们仅在用户实际发出语音时，利用口腔运动进行验证。我们借助卷积神经网络取得了重大进展，并通过实际无声语音与一台未经改装的智能音箱（Amazon Alexa）进行交互，完成了概念验证。

本文介绍了一种仅基于超声图像的无声语音交互系统，名为“SottoVoce”（图 1）。通过结合两种深度神经网络，该系统能够经训练后，根据超声成像设备采集的图像序列生成语音信号。我们的贡献可概括为以下两个方面：

- 本文提出了一种双层深度卷积神经网络模型，用于将超声图像转换为真实声音。
- 作为概念验证，我们开发了一套无声语音系统，并证明该系统能够控制语音控制设备（本例中为 Amazon Echo），且无需对设备进行改装。

2 相关工作

2.1 无声语音

研究人员使用各种技术和方法研究了无声语音接口 [8, 31]。唇读 [52] 或面部图像 [12] 可用于估计受试者在不使用音频信息的情况下说出的语音。

SilentVoice [17] 是一种吸气语音 方法，能够捕捉极其轻柔的语音。电磁发音运动描记术（EMA）已用于开发脑机接口 [5]及其他交互式应用 [53]。可将磁体贴在受试者身上，以检测无声语音 [14, 19, 23]。脑电图（EEG） [44]和肌电图（EMG） [37, 49] 也是无声语音识别的典型方法。尤其是，EMG 已应用于交互用途以及人机交互（HCI）相关应用 [38]，例如控制网页浏览器 [32]。Kapur 等人最近提出的一个方案 [33] 使用多个电极，感测受试者内在语音中的神经肌肉信号。除上述方法外，研究人员还结合并测量了其他人脸电位 [22, 48]。

超声成像也被用于语音处理领域中的无声语音识别 [9, 27]。在一项专注于歌唱的研究中，研究人员基于超声波和唇部视频合成了歌唱元音 [30]。在一项类似研究中，研究人员利用超声波以及舌头和嘴唇的光学图像，开发了一种无声语音界面 [26, 28]。在另一项研究中，研究人员创建了一种映射技术，可根据原始超声图像自动生成舌头运动的动画 [13]。

其他方法将上述技术与深度神经网络相结合，例如脑机接口应用 [4], 、肌电信号 [10], 以及利用长短期记忆网络（ LSTM）进行唇读 [52]。将深度神经网络与超声成像相结合，用于无声语音界面的方法也已受到关注。例如，研究人员利用超声成像捕捉舌头运动，并将其与深度神经网络相结合 [7]。此外，研究人员还利用深度神经网络估计了此前被认为无法预测的基频（F0）曲线 [21]。另外，有研究提出，一种称为“ Eigentongues”的全局视听建模方法，在使用神经网络时的表现优于舌部轮廓建模 [25]。研究人员还提出了一种基于深度神经网络的无声语音研究新基准 [31]。这些方法仍处于基础概念验证阶段，尚未针对控制现有的语音交互设备（如智能音箱）进行评估。

基于上述介绍的先前方法，我们旨在开发一种系统，使无声语音能够与语音控制设备进行交互，并在 HCI 环境中用于交互目的。此外，为了提升所开发系统的性能，我们还致力于通过人类与人工神经网络之间的交互来提升其性能。

2.2 非听觉输入与交互

EarFieldSensing [39] 是一种基于电场感测的手势识别技术。CanalSense [3] 可感测面部移动时耳道内气压的变化。Tongue-in-Cheek [18] 利用 X 波段多普勒雷达感测舌头的运动，以识别面部手势。其他使用肌电图 [55] 或结合脑部与肌肉信号感测 [43] 的方法也值得关注。

上述技术已用于手臂手势识别；不过，已有研究展示了一种利用超声成像实现这一目的的方法，即 EchoFlex [40]。它是一种交互传感器，利用超声成像识别前臂肌肉的运动。该研究结果表明，这种传感器表现良好，并有望取代其他现有方法。这些结果也启发作者考虑利用口腔内部的超声成像实现无声语音交互。

2.3 与智能设备交互

得益于移动设备和智能设备的发展，我们如今可以通过各种方式与它们频繁交互 [35]。我们可以使用语音作为输入方式 [47],，也已有研究提出了克服此类语音控制用户界面相关问题的方法 [41]。一种典型的界面（下文称为个人“代理”）是 Amazon Echo，通常被称为“Alexa”，人们可以与之交流和聊天。人们对这种代理的影响进行了广泛研究，结果表明，它是一种能够满足人们需求或影响我们生活的有效代理 [36, 45, 50]。

人工智能（人工智能）在个人智能助理（如 Alexa）以外的任务中，作用日益重要；如今，我们会与人工智能互动和协作，例如发送短信 [24] 以及设计物品 [16]。人类与人工智能共存并互动是一个引人瞩目的议题，它能够应对社会问题并改善我们的生活质量。Glove Talk II [15], 于1995年推出，是一种手势转语音系统，它利用神经网络将手势转换为语音合成器的10个控制参数。然而，使用该系统需要用户接受约100小时的长期训练。本研究参考这项工作，应用近期改进的深度神经网络技术，打造人类与人工智能之间的混合式互动，让用户（人类）学习如何适应并使用人工智能系统，以实现更好的互动。

3 SottoVoce 系统架构

图 2 展示了所提出的、用于从超声图像生成声音的系统架构。总体而言，该系统旨在将一种序列表示（此处为超声图像）转换为另一种序列表示（此处为语音）。这一目标与文本转语音系统 [54], 语音转换系统 [2], 以及唇读或面部到语音系统 [12] 类似。受这些系统启发，所提出的系统采用了两个神经网络。

第一个神经网络（图 2 中的“网络 1”）将超声图像时间序列转换为声音表示向量。我们使用了一个 64 维的梅尔尺度频谱，频率范围为 300 Hz 至 8，000 Hz，每隔 20 ms 采样一次，并将其作为声音表示向量。随后，转换得到的声音表示构成一系列声音表示向量（即频谱图）。该声音频谱图可以转换为音频信号。此外，为了提高这些向量的质量，还会通过下一个神经网络（图 2 中的“网络 2”）对其进行转换，该网络会生成一系列与输入声音表示向量长度相同的声音表示向量。最后，输出向量会转换为实际的音频信号。

这两个网络均为说话人相关；因此，为了训练它们，系统需要一组用户说出各种语音命令时采集的超声成像视频。

3.1 超声成像设备

超声成像设备采用 CONTEC CMS600P2 全数字 B 超诊断系统。用户将 3.5兆赫凸阵超声成像探头贴于下颌下方（图 3）。该系统提供一个屏幕输出端口，可连接至显示器。此外，还使用了显示数字化单元，将发送到显示器的信号转换为 MPEG-4 电影文件。图 4 展示了获取的超声图像。

我们发现超声图像与声音采集之间存在延迟。为补偿这一延迟，我们检查了视频中对应的发音和舌头运动，并估算出延迟为 300 ms。随后，我们在训练数据中对这一延迟进行了调整。

3.2 网络 1

网络 1 使用一系列 K 超声图像（尺寸为 128×128，单色）作为输入，并生成 n维声音表示（梅尔尺度频谱）作为输出。目前，使用 K 为 13、 n 为 64。由于超声图像的帧率为每秒 30 帧，因此 K 超声图像的时长

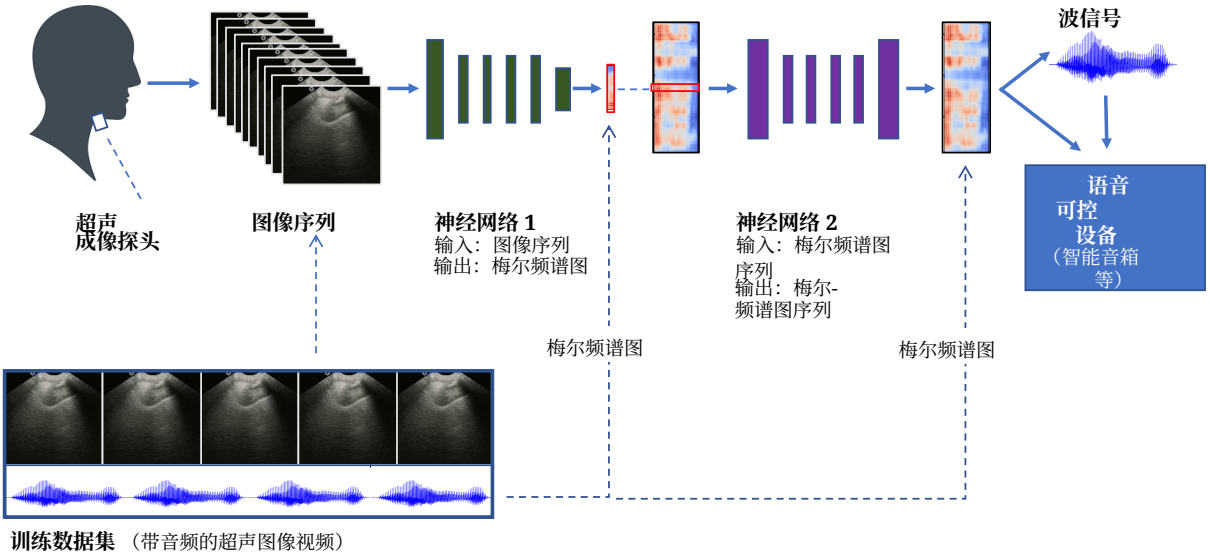

图 2：SottoVoce 系统概览



图 3：超声成像探头

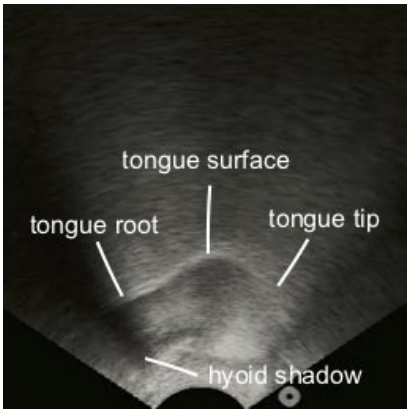

图 4：从下方固定在下颌上的探头获取的超声图像。

为 400 毫秒。这段时长涵盖了话语的静态和运动特征。超声图像样本如图 5 所示。重复准备 K 张图像组成的序列，使每 20 毫秒生成一个梅尔尺度频谱（即帧移为 20 毫秒）（图 6）。

从音频信号数据中提取与每个超声图像序列中心时刻对应的声音表示向量，并训练网络 1 来生成该向量。

网络 1 基于卷积神经网络（CNN），由四层组成：二维卷积 - 带泄漏的修正线性单元 - 随机失活 - 批归一化，随后是六层：展平 - 全连接层 - 带泄漏的修正线性单元 - 随机失活 - 全连接层 - 带泄漏的修正线性单元。网络 1 的输出大小与声音表示向量的长度相同（即 64）。输入图像和输出向量分别归一化至 0 到 1。损失函数为均方误差，优化器为 Adam [34]。

3.3 网络 2

为提高声音质量，网络 2 使用一组声音表示向量，并生成与输入等长的一组声音表示向量。该模型由一组一维卷积滤波器（一维卷积）组成，卷积核大小从 1 到 M（当前使用$M = 8$），后接 U形网络 [46]；该网络包含三层 一维卷积 - 最大池化（步幅=2）- 带泄漏的修正线性单元 - 随机失活，以及三组 一维反卷积 - 拼接（图 7）。一维卷积滤波器组对输入序列的局部信息和上下文信息进行显式建模。随后，U形网络通过精确定位进一步改善音频序列的质量。最后，网络生成梅尔频谱向量。

为了训练网络 2，使用网络 1 从训练用超声视频片段的图像中生成梅尔频谱向量作为输入，并将同一训练视频片段音频中长度相同的梅尔频谱向量作为输出。与网络 1 的情况类似，使用均方误差作为损失函数，并使用 Adam 作为优化器。

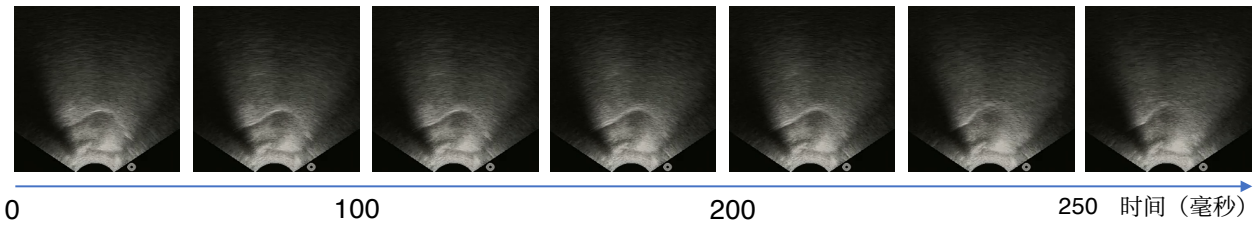

图 5：受试者即将发出“Alexa”时，其喉部的一系列超声图像。

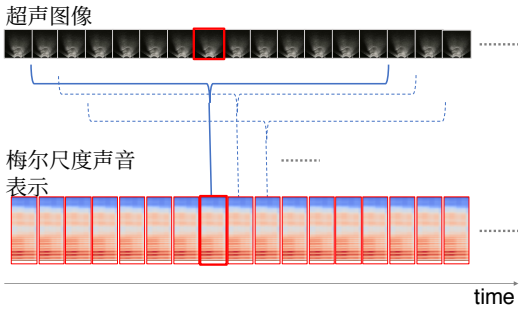

图 6：训练数据的表示。K 个超声图像与相应的声音（梅尔尺度声音向量）配对。

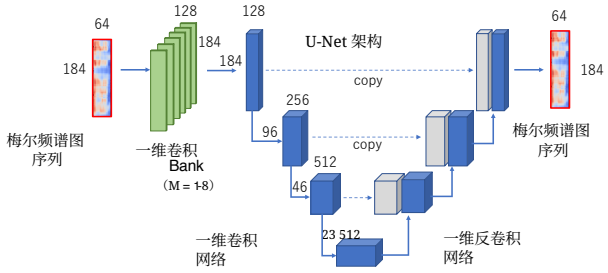

图 7：网络 2 提高了生成的梅尔尺度频谱序列的质量。（注：为与神经网络示意图保持一致，音频特征向量的时间轴显示为垂直轴。）

为简化起见，输入和输出的时间长度固定为相同值（目前使用 3.68 *s*）。此时长涵盖了许多常见的语音命令。

3.4 声音生成

经过神经网络处理后，一系列梅尔尺度频谱声音表示向量通过 Griffin Lim 算法 [20] 转换为音频信号。此转换可以基于网络 1 或网络 2 的输出完成。

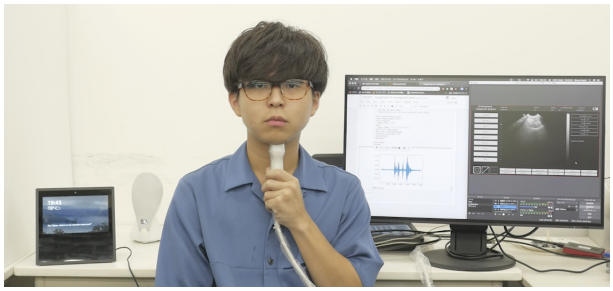

图 8：用于训练和评估的设备。

在测试中，生成的音频信号通过音频扬声器播放，并可用于控制附近的声音控制设备，例如智能音箱。我们也在考虑直接将音频波形信号作为语音控制设备的音频输入信息，而不实际将其还原为声波。

3.5 训练

在准备训练数据时，两名协作者（一名 28 岁男性和一名 24 岁男性）将超声成像探头固定在下颌下方，并按照指示发出各种语音命令。每位协作者约贡献了 500 条语音命令（表 ??）。对于每条命令，除了录制语音话语外，还录制了超声图像视频。训练过程经作者所在机构的研究伦理委员会批准。

录制的视频用于训练网络 1。超声图像被缩放至 128×128，并用作输入。相应的语音话语被转换为梅尔尺度频谱，并用作输出。

网络 2 的测试集数量与训练用视频文件的数量相同（约 500 个）。为增加测试集数量，采用了数据增强方法，即对输入的梅尔频谱向量添加高斯噪声。

由于我们的模型与说话人相关，因此针对每位说话人分别训练网络 1 和网络 2。先训练网络 1，随后用它创建用于训练网络 2 的数据集。

3.6 实现细节

上述网络模型基于 Keras [6] 深度学习平台实现，并以 Tensorflow [1] 作为

后端，采用 NVIDIA GeForce 1080ti 作为 GPU 板卡。使用 500 条语音命令训练网络 1（为网络 1 创建 35,000 组成对的训练数据），耗时约 4 小时。训练网络 2 不到一小时。注释：由于超声成像设备无法直接连接到运行神经网络的 Ubuntu 机器，因此开发了一个简单的服务器—客户端程序。它将控制超声成像设备的计算机与运行神经网络的计算机连接起来。生成的音频信号会被发送回连接超声成像设备的计算机。

神经网络处理超声图像时，需要以 3.68 s的时长/间隔处理这些图像。总处理时间（包括视频处理、神经网络处理，以及将梅尔尺度频谱转换为音频波形）为 2.36s 。

4 结果

将超声图像转换为声音的结果如图9所示。图中，顶部一行的图表显示声音表示向量（梅尔刻度频谱图），底部一行的图表则为相应的波形。标记为 Net1 的图表是网络 1 的结果，标记为 Net2 的图表是网络 1 + 网络 2 的结果，而标记为“original”的图表则是编码为梅尔刻度频谱图后再解码回波形的音频数据。因此，最后一个被视为训练的真实值。尽管网络 1 和网络 2 的输出之间的差异并不明显，但我们观察到，网络 2 生成的声音优于网络 1 生成的声音（输出音频信号的示例见补充视频。）

值得注意的是，网络 1 和网络 2 都生成了自然语调，这通常被认为是由声带振动而非内部状况引起的。这一结果或许表明，神经网络学会了语音的上下文。

随后，我们使用一台现有的（未作改动的）智能音箱（Amazon Echo 和 Amazon Echo Show）对计算机扬声器发出的生成的声音进行了测试，结果证实，生成的声音能够控制智能音箱。用于训练和测试的语音命令都是典型的 Amazon Alexa 命令。在这项测试中，参与者将以下四条命令各说五遍（共 20 句话）：“Alexa，播放音乐”、“Alexa，天气怎么样”、“Alexa，现在几点了”和“Alexa，播放爵士乐”。

表 1 列出了网络 1、网络 1 + 网络 2 以及原始版本（经梅尔尺度编码和解码后）作为真实值时的识别成功率。我们确认，网络 1 与网络 2 的组合提高了识别率。我们还注意到，触发词（“Alexa”）总是能被清晰地重新生成。这可能是因为该词只是训练集中发音最突出的词。

我们还使用 Google 的云端语音转文本引擎 [29]，并在与智能音箱识别率测量相同的环境下进行了词错误率（WER）测量。GT、网络 1 和网络 2 的词错误率分别为 20.61%、41.03% 和 33.56%（数据取自表 1 中两位用户发出的共 40 条语音命令的平均值）。

	用户 A	用户 B	ave.
网络 1	60.0%	25.0%	42.5%
网络 1 + 网络 2	65.0%	65.0%	65.0%
GT	90.0%	90.0%	90.0%

表1：在未更改的现有智能音箱上进行测试的语音识别成功率（“GT”表示由原始语音数据经梅尔尺度编码/解码得到的结果，视为训练的真实值）。

我们认为，这也能证明网络 2 的有效性。通过这些研究，我们发现，诸如“今天天气怎么样？”之类的命令在所有条件下都具有较高的识别率，而“播放爵士乐”等较短的命令表现则差得多。这或许表明，较长的命令更便于网络 2 获取上下文信息。

5 端到端评估与观察

我们检验了真实场景下端到端的无声语音到音频转换。在这种情况下，用户需要做出口腔动作，说出语音命令，但不实际发出声音；超声成像探头会记录口腔运动。随后，所提出的系统将获得的图像序列转换为语音。

我们要求参与者尽可能无声地说话（不振动声带），并尽量使无声说话时的口腔动作与发声时相似。不过，我们没有要求他们屏住呼吸。因此，偶尔会听到微弱的漏音。为说明这一情况，我们按照 SilentVoice [17] 的评估方法测量了参与者发出的声音强度。我们使用了一台规格相同的噪声计（最小量程 30 dB, 1.5 dB 误差），将其放置在 30 cm 处；房间的背景噪声级为 31.0 dB（A）。20 次测量（典型的 Alexa 语音命令）所得峰值声音强度的平均值为 37.14 dB（A），低于轻声耳语的音量。

图10展示了用户发出语音和未发出语音时生成的声音对比（实际生成的声音请参阅补充视频）。

起初，结果并不理想。我们最初预期，用户未发出语音时口腔内的运动（图10（中））与用户实际发出语音时的运动（图10（左））相同；然而，我们发现两者之间存在细微差异，而且无语音图像生成的声音质量不如有语音图像生成的声音。

然而，我们观察到了以下有趣的现象。由于用户也能听到由无语音图像生成的声音，用户便尝试调整口腔运动，以获得稍好的结果。经过多次尝试，生成声音的质量有所提升。在这种情况下，我们认为用户提升了自身的无声发声技巧。

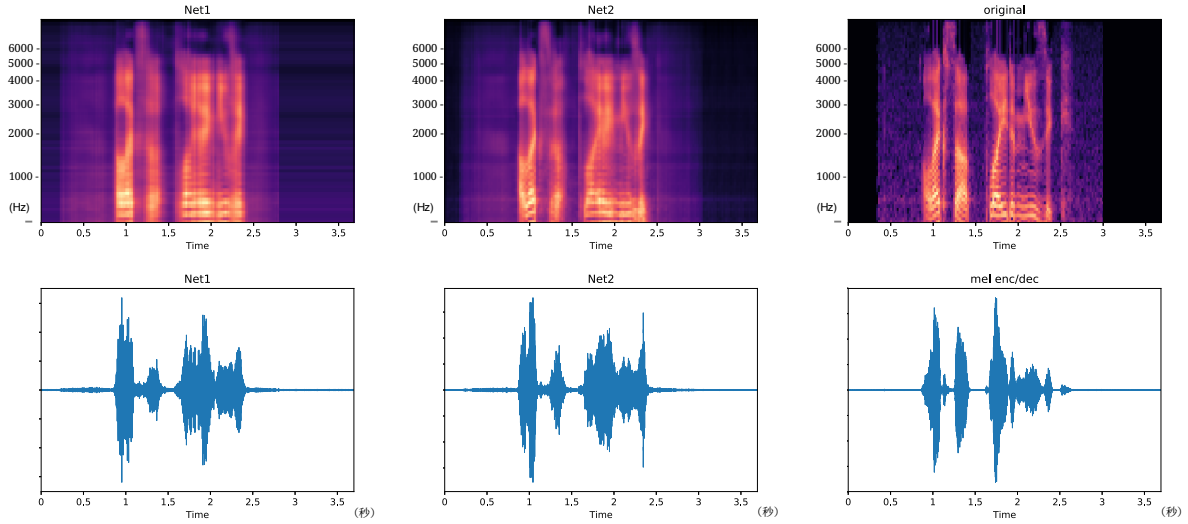

图9：训练结果（以波形和梅尔刻度频谱图表示）：网络1：网络 1 的结果；网络2：网络 2 的结果；原始：通过梅尔刻度频谱图编码并解码为音频信号的原始语音。这是本次训练的“真实值”。

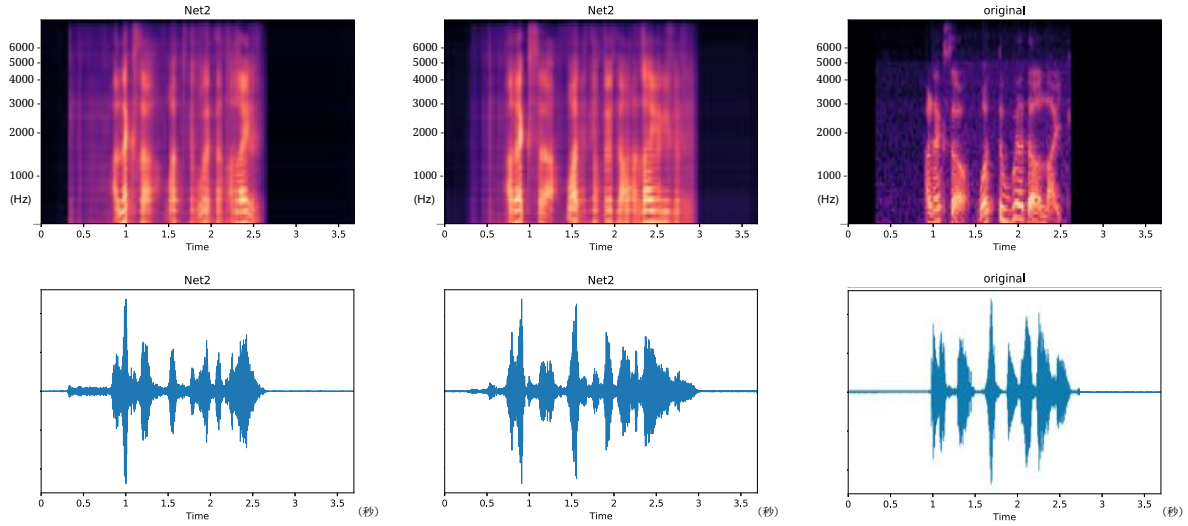

图10：生成的声音对比。左：用户发出语音时，网络 2 根据超声图像生成的结果；中：用户未发出语音时，网络 2 根据超声图像生成的结果；右：原始语音（经梅尔刻度频谱图编码/解码）

6 讨论

6.1 渐进式语音生成

在当前设计中，我们以语音命令为粒度（约 3.6 秒），将获取的超声图像序列转换为语音。这是因为网络 2 使用固定长度的语音表示序列。然而，根据对用户练习情况的观察，逐步生成声音会更好，这样用户便能获得更紧密的反馈回路，从而学习如何通过口腔运动生成更好的语音。

6.2 向人体持续发射超声波

将超声波持续发射到人体内对人体器官的影响尚不明确。不过，我们或许可以结合一种简单的触发机制，来启动和停止超声波的发射。例如，设备中的加速度计和麦克风相结合，能够检测下颌运动，从而在不实际发出声音的情况下启动（无声）语音命令。



图 11：未来图景：超声探头与开放式耳机的可能配置

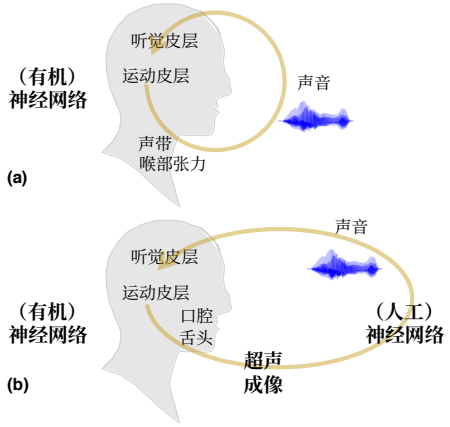

图 12：人机融合：人工神经网络与有机神经网络形成反馈回路，以获得更好的结果。

6.3 声带障碍的治疗

我们也希望声带受损的人能够使用我们的研究成果。正如在人机融合一节中所述，即使声带无法正常工作，人们仍能学会如何正确控制口腔和舌头来发出声音。

6.4 与其他模态结合

最后需要指出的是，本研究并非旨在排斥其他模态。结合肌电图、加速度计和NAM麦克风的信息，可能会提升语音识别的质量。研究这些模态的组合是未来的研究方向。

6.5 人机融合

上述观察表明，人类与人工智能之间存在一种有趣的关系。与其将人工智能视为一个自主或独立的实体，我们或许可以将其看作人类的一部分。因此，即使（人工）神经网络的初始性能并不完美，用户也能逐步学习并提升其性能。

这类似于人们学习基本技能的方式。当我们学习话语时，驱动口腔、舌头和声带的运动皮层与听觉皮层之间的协调会形成一个紧密回路，从而获得更好的语音表现（图12（a））。通过扩展这一回路，有机神经网络（例如我们的大脑）与人工神经网络也可以形成一个紧密的反馈回路（图12（b））。我们将这种形成称为“人机融合”，而非人机交互。

在这方面，我们提及1995年的Glove Talk II研究——这是人机融合领域的一项开创性工作 [15]。在这项研究中，用户学会了用手势控制语音合成器。研究中使用了三个简单的神经网络，而这位用户（当时是一名钢琴家）花了100多个小时才生成可听见的语音。采用更好的神经网络和学习器相结合，或许可以缩短这一学习时间。

7 结论

本文提出了一种基于超声成像的无声语音交互方法。该方法依次使用两个神经网络，将用户在没有发声的情况下以口腔动作表达的“话语”转换为声音（语音），并可用于操作现有的语音可控设备，例如智能音箱。

基于这一结果，我们设想，未来可穿戴计算机的形态将结合一枚可贴附于下颌下方的超声成像探头，以及一副骨传导耳机或开放式耳机（图 11）。采用这种配置后，用户无需发出声音，便可随时调用语音控制助手并获取响应。

参考文献

[1] Martín Abadi、Ashish Agarwal、Paul Barham、Eugene Brevdo、Zhifeng Chen、Craig Citro、Greg S. Corrado、Andy Davis、Jeffrey Dean、Matthieu Devin、Sanjay Ghemawat、Ian Goodfellow、Andrew Harp、Geoffrey Irving、Michael Isard、Yangqing Jia、Rafal Jozefowicz、Lukasz Kaiser、Manjunath Kudlur、Josh Levenberg、Dandelion Mané、Rajat Monga、Sherry Moore、Derek Murray、Chris Olah、Mike Schuster、Jonathon Shlens、Benoit Steiner、Ilya Sutskever、Kunal Talwar、Paul Tucker、Vincent Vanhoucke、Vijay Vasudevan、Fernanda Viégas、Oriol Vinyals、Pete Warden、Martin Wattenberg、Martin Wicke、Yuan Yu 和 Xiaoqiang Zheng。2015年。《TensorFlow：异构系统上的大规模机器学习》。https://www.tensorflow.org/ 软件可从 tensorflow.org 获取。

[2] Dabi Ahn。2017。使用非平行数据进行语音转换。https://github.com/andabi/deep-voice-conversion。[3] Toshiyuki Ando、Yuki Kubo、Buntarou Shizuki 和 Shin Takahashi。2017。CanalSense：基于耳道气压感测的面部相关动作识别系统。载于*第30届ACM用户界面软件与技术年会论文集（加拿大魁北克省魁北克市）（UIST '17）*。ACM，纽约州纽约市，美国，679–689。https://doi.org/10.1145/3126594.3126649。[4] 2014。/。F Bocquelet、Thomas Hueber、Laurent Girin、Pierre Badin 和 B Yvert。。面向脑机接口应用的深度神经网络鲁棒发音语音合成。，2288-2292页。[5] Florent Bocquelet、Thomas Hueber、Laurent Girin、Christophe Savariaux 和 Blaise Yvert。2016。面向脑机接口的基于发音的语音合成器实时控制。*PLOS计算生物学* 12，11（2016年11月），1–28。https://doi.org/10.1371/journal.pcbi.1005119。[6] 2015。/。François Chollet 等。。Keras。https://keras.io。[7] Tamás Gábor Csapó、Tamás Grósz、Gábor Gosztolya、László Tóth 和 Alexandra Markó。2017。面向无声语音接口的基于深度神经网络的超声波到语音转换。载于*INTERSPEECH*。[8] B. Denby、T. Schultz、K. Honda、T. Hueber、J. M. Gilbert 和 J. S. Brumberg。2010。无声语音接口。*Speech Commun.* 52，4（2010年4月），270–287。https://doi.org/10.1016/j.specom.2009.08.002。[9] 2004。。利用舌头实时超声图像进行语音合成。载于*2004年IEEE声学、语音与信号处理国际会议*，第1卷，685–688。[10] 2015。I–。https://doi.org/10.1109/ICASSP.2015.7178064。。L. Diener、M. Janke 和 T. Schultz。。使用深度神经网络将面部肌电信号直接转换为语音。载于*2015年国际联合*

神经网络会议（IJCNN）。1–7。https://doi.org/10.1109/IJCNN.2015. 7280404[11] Roman Gr. Maev（编）。2013。《声学显微镜与高分辨率成像进展：从原理到应用》。Wiley-VCH。[12] Ariel Ephrat、Tavi Halperin 和 Shmuel Peleg。2017。基于无声视频改进语音重建。ICCV 2017视听媒体计算机视觉研讨会（2017）。[13] Diandra Fabre、Thomas Hueber、Laurent Girin、Xavier Alameda-Pineda 和 Pierre Badin。2017。根据声道超声图像自动驱动发音舌模型动画。Speech Communication 93（2017），63–75。https://doi.org/ 10 1016 2017 08 002 。/j.specom。。。M.J. Fagan、S.R. Ell、J.M. Gilbert、E. Sarrazin 和 P.M. Chapman。2008。为喉切除术后的患者开发（无声）语音识别系统。Medical Engineering & Physics 30，4（2008），419–425。https://doi.org/ 10 1016 2007 05 003 。/j.medengphy。。。[14] Sidney Fels 和 Geoffrey Hinton。。Glove-TalkII：一种自适应的手势到共振峰接口。载于SIGCHI计算机系统中的人为因素会议论文集（美国科罗拉多州丹佛）（CHI ’95）。ACM Press/Addison-Wesley Publishing Co.，纽约州纽约市，美国，456–463。https://doi.org/10.1145/223904.223966[15] Rebecca Fiebrink。2017。机器学习作为元乐器：塑造表现性器乐创作的人机合作。Springer Singapore，新加坡，1995 137 151 10 1007 978 981 10 2951 6_10 – 。https://doi.org/ 。/ - - - -[16] Masaaki Fukumoto。。SilentVoice：通过吸气语音实现不易察觉的语音输入。载于第 31 届ACM用户界面软件与技术年会论文集（德国柏林）（UIST ’18）。ACM，纽约州纽约市，美国，2018 237 246 10 1145 3242587 3242603 – 。https://doi.org/ 。/ 。[17] Mayank Goel、Chen Zhao、Ruth Vinisha 和 Shwetak N. Patel。2015。Tongue-in-Cheek：利用无线信号实现非侵入式、灵活的面部手势检测。载于第33届ACM计算机系统中的人为因素会议论文集（韩国首尔）（CHI ’15）。ACM，纽约州纽约市，美国，255 258 10 1145 2702123 2702591 – 。https://doi.org/ 。/ 。[18] Jose A. Gonzalez、Lam A. Cheah、James M. Gilbert、Jie Bai、Stephen R. Ell、Phil D. Green 和 Roger K. Moore。2016。基于永磁体发音测量与直接合成的无声语音系统。Comput. Speech Lang. 39，C（2016年9月 2016 67 87 10 1016 2016 02 002 ），– 。https://doi.org/ 。/j.csl。。。[19] D. Griffin 和 Jae Lim。。基于修改后的短时傅里叶变换进行信号估计。IEEE声学、语音与信号处理汇刊 32，2（4月 1984 236 243 10 1109 1984 1164317 ），– 。[20] https://doi.org/ 。/TASSP。。Tamás Grósz、Gábor Gosztolya、László Tóth、Tamás Csapó 和 Alexandra Markó。2018。面向基于深度神经网络的超声无声语音接口的基频估计。[22] Tatsuya Hirahara、Makoto Otani、Shota Shimizu、Tomoki Toda、Keigo Nakamura、Yoshitaka Nakajima 和 Kiyohiro Shikano。2010。利用体传导声道共振信号增强无声语音。Speech Communication 52，4 2010 301 313 10 1016 2009 12 001（），– 。https://doi.org/ 。/j.specom。。。无声语音接口。[23] Robin Hofe 、 Stephen R. Ell 、Michael J. Fagan、James M. Gilbert、Phil D. Green、Roger K. Moore 和 Sergey I. Rybchenko。2013。基于磁感测的无声语音接口的小词汇量语音识别。Speech Commun. 55, 1 2013 22 32 10 1016 2012 02 001 （1月），– 。[24] https://doi.org/ 。/j.specom。。。Jess Hohenstein 和 Malte Jung。。人工智能辅助消息传递：对人工智能支持下人人文本对话的研究。载于2018年CHI计算机系统中的人为因素会议扩展摘要（加拿大魁北克省蒙特利尔）（CHI EA ’18）。ACM，纽约州纽约市，美国，文章编号LBW089，共6页。https://doi.org/ 10 1145 3170427 3188487 。/ 。[25] T. Hueber、G. Aversano、G. Cholle、B. Denby、G. Dreyfus、Y. Oussar、P. Roussel 和 M. Stone。2007。面向基于超声波的无声语音接口的特征舌特征提取。载于2007年IEEE声学、语音与信号处理国际会议（ICASSP ’07），第1卷。I–1245–I–1248。https://doi.org/10.1109/ICASSP.2007.366140[26] Thomas Hueber、Elie-Laurent Benaroya、Gérard Chollet、Bruce Denby、Gérard Dreyfus 和 Maureen Stone。2010。由舌和嘴唇的超声图像与光学图像驱动的无声语音接口的开发。Speech Commun. 52, 4 2010 288 300 10 1016 2009 11 004 （4月），– 。[27] https://doi.org/ 。/j.specom。。。Thomas Hueber、Elie-Laurent Benaroya、Bruce Denby 和 Gérard Chollet。2011。面向基于超声波的无声语音接口的发音数据与声学数据之间的统计映射。载于INTERSPEECH。[28] Thomas Hueber、Gerard Chollet、Bruce Denby 和 M Stone。2008。为无声语音接口应用采集超声波、视频和语音声学数据。（2008年1月）。[29] Google公司 [不详]。Clund语音转文本。https://cloud.google.com/speech-to-text/.[30] Aurore Jaumard-Hakoun、Kele Xu、Clémence Leboullenger、Pierre Roussel-Ragot 和 Bruce Denby。2016。利用舌和嘴唇成像实现基于发音的歌唱语音合成。载于ISCA Interspeech 2016，第2016卷。美国旧金山，1467 1471 10 21437 2016 385 – 。[31] https://doi.org/ 。/Interspeech。- Yan Ji、Licheng Liu、Hongcui Wang、Zhilei Liu、Zhibin Niu 和 Bruce Denby。2018。通过深度学习更新无声语音挑战赛基准。Speech Commun. 98, 2018 42 50 10 1016 2018 02 002 C（4月），–。[32] https://doi.org/。/j.specom.。。Chuck Jorgensen 和 Kim Binsted。。基于肌电图的次声语音识别网页浏览器控制。载于《第38届夏威夷国际系统科学会议论文集》第09卷（HICSS ’05）。IEEE计算机学会，美国华盛顿特区，294.3–。https://doi.org/10.1109/HICSS.2005.683[33] Arnav Kapur、Shreyas Kapur 和 Pattie Maes。2018。AlterEgo：个性化可穿戴无声语音接口。载于第23届智能用户界面国际会议（日本东京）（IUI ’18）。ACM，美国纽约州纽约，43–53。https: //doi.org/ 10 1145 3172944 3172977 . / 。Diederik P. Kingma 和 Jimmy Ba。2014。Adam：一种随机优化方法。CoRR abs/1412.6980（2014）。arXiv:1412.6980 http://arxiv.org/abs/1412.6980[35] Lorenz Cuno Klopfenstein、Saverio Delpriori、Silvia Malatini 和 Alessandro Bogliolo。2017。机器人崛起：对话式界面、模式与范式综述。载于《2017年交互系统设计会议论文集》（英国爱丁堡）（DIS ’17）。ACM，美国纽约州纽约，555– 565 10 1145 3064663 3064672 。https://doi.org/。/。Rafal Kocielnik、Daniel Avrahami、Jennifer Marlow、Di Lu 和 Gary Hsieh。2018。面向工作场所反思的设计：基于聊天和语音的对话代理。载于《2018年交互系统设计会议论文集》（中国香港）（DIS ’18）。ACM，美国纽约州纽约，881–894。https://doi.org/10. 1145/3196709.3196784[37] L. Maier-Hein、F. Metze、T. Schultz 和 A. Waibel。2005。使用表面肌电图进行与会话无关的非听觉语音识别。载于《2005年IEEE自动语音识别与理解研讨会》。331–336。https://doi.org/ 10 1109 2005 1566521 . /ASRU.。Hiroyuki Manabe、Akira Hiraiwa 和 Toshiaki Sugimura。2003。“使用肌电图进行清音语音识别——哑语语音识别”。载于《CHI ’03计算机系统中的人为因素会议扩展摘要》（美国佛罗里达州劳德代尔堡）（CHI EA ’03）。ACM，美国纽约州纽约，794–795。https://doi.org/10.1145/765891. 765996[39] Denys J. C. Matthies、Bernhard A. Strecker 和 Bodo Urban。2017。EarField-Sensing：一种新型耳内电场感测技术，通过面部表情丰富可穿戴手势输入。载于《2017年CHI计算机系统中的人为因素会议论文集》（美国科罗拉多州丹佛）（CHI ’17）。ACM，美国纽约州纽约，1911 1922 10 1145 3025453 3025692 –。https://doi.org/。/。[40] Jess McIntosh、Asier Marzo、Mike Fraser 和 Carol Phillips。2017。EchoFlex：使用超声成像进行手势识别。载于《2017年CHI计算机系统中的人为因素会议论文集》（美国科罗拉多州丹佛）（CHI ’17）。ACM，美国纽约州纽约，1923–1934。https://doi.org/10.1145/3025453. 3025807[41] Chelsea Myers、Anushay Furqan、Jessica Nebolsky、Karina Caro 和 Jichen Zhu。2018。语音用户界面中用户克服障碍的模式。载于《2018年CHI计算机系统中的人为因素会议论文集》（加拿大魁北克省蒙特利尔）（CHI ’18）。ACM，美国纽约州纽约，文章6，7页。https://doi.org/ 10 1145 3173574 3173580 . /。Y. Nakajima、H. Kashioka、K. Shikano 和 N. Campbell。2003。使用贴附于皮肤的听诊器式麦克风进行非听觉喃喃语识别输入界面。载于《2003年IEEE声学、语音与信号处理国际会议论文集》（ICASSP ’03），第5卷。V–708。https://doi.org/10.1109/ ICASSP.2003.1200069[43] Phuc Nguyen、Nam Bui、Anh Nguyen、Hoang Truong、Abhijit Suresh、Matt Whitlock、Duy Pham、Thang Dinh 和 Tam Vu。2018。TYTH——在你的牙齿上打字：用于人机界面的舌齿定位。载于《第16届年度国际移动系统、应用与服务会议论文集》（德国慕尼黑）（MobiSys ’18）。ACM，美国纽约州纽约，269–282。https://doi.org/ 10 1145 3210240 3210322 . /。Anne Porbadnigk、Marek Wester、Jan Calliess 和 Tanja Schultz。2009。基于脑电图的语音识别——时间效应的影响。载于 BIOSIGNALS。[45] Amanda Purington、Jessie G. Taft、Shruti Sannon、Natalya N. Bazarova 和 Samuel Hardman Taylor。2017。“Alexa是我的新闺蜜”：Amazon Echo的社会角色、用户满意度与拟人化。载于《2017年CHI计算机系统中的人为因素会议扩展摘要》（美国科罗拉多州丹佛）（CHI EA ’17）。ACM，美国纽约州纽约，2853–2859。https://doi.org/ 10 1145 3027063 3053246 . /。Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。2015。U-Net：用于生物医学图像分割的卷积网络。CoRR abs/1505.04597（2015）。arXiv: 1505 04597 1505 04597 。http://arxiv.org/abs/。Alexander I. Rudnicky。。语音驱动界面的设计。载于《语音与自然语言研讨会论文集》（美国宾夕法尼亚州费城）（HLT ’89）。计算语言学协会，美国宾夕法尼亚州斯特劳兹堡，120–124。https://doi.org/10.3115/100964.100972[48] Himanshu Sahni、Abdelkareem Bedri、Gabriel Reyes、Pavleen Thukral、Zehua Guo、Thad Starner 和 Maysam Ghovanloo。2014。舌耳接口：用于无声语音识别的可穿戴系统。载于《2014年ACM可穿戴计算机国际研讨会论文集》（美国华盛顿州西雅图）（ISWC ’14）。ACM，美国纽约州纽约，47–54。https://doi.org/10.1145/2634317.2634322

[49] Tanja Schultz。2010。ICCHP主题演讲：基于肌电图识别无声语音和弱语音。载于帮助特殊需求人士的计算机，Klaus Miesenberger、Joachim Klaus、Wolfgang Zagler 和 Arthur Karshmer（编）。Springer Berlin Heidelberg，德国柏林海德堡，595–604。[50] Alex Sciuto、Arnita Saini、Jodi Forlizzi 和 Jason I. Hong。2018。“嘿，Alexa，最近怎么样？”：对家庭对话代理使用情况的混合方法研究。载于《*2018*年交互系统设计会议论文集》（中国香港）（*DIS '18*）。ACM，美国纽约州纽约，857–868。https://doi.org/10.1145/ 3196709.3196772[51] László Tóth、Gábor Gosztolya、Tamás Grósz、Alexandra Markó 和 Tamás Csapó。2018。面向基于超声波的无声语音接口的语音识别与语音合成参数多任务学习。[52] Michael Wand、Jan Koutník 和 Jürgen Schmidhuber。2016。基于长短期记忆的唇读。*CoRR* abs/1601.08188（2016）。arXiv:1601.08188 http://arxiv.org/abs/ 1601 08188 . Jun Wang、Ashok Samal 和 Jordan Green。2014。《基于电磁发音仪的实时交互式无声语音接口初步测试》。（2014年6月）。[53] [54] Yuxuan Wang、R. J. Skerry-Ryan、Daisy Stanton、Yonghui Wu、Ron J. Weiss、Navdeep Jaitly、Zongheng Yang、Ying Xiao、Zhifeng Chen、Samy Bengio、Quoc V. Le、Yannis Agiomyrgiannakis、Rob Clark 和 Rif A. Saurous。2017。《Tacotron：一种完全端到端的文本转语音合成模型》。*CoRR* abs/1703.10135（2017）。arXiv: 1703 10135 1703 10135 。http://arxiv.org/abs/ 。Qiao Zhang、[55] Shyamnath Gollakota、Ben Taskar 和 Raj P.N. Rao。2014。《非侵入式舌头机器接口》。载于*SIGCHI*计算机系统中的人为因素会议论文集（加拿大安大略省多伦多）*(CHI '14)*。ACM，纽约，NY，美国，2555–2558。https://doi.org/10.1145/2556288.2556981