УДК 004.934.2

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ ВОВЛЕЧЕННОСТИ ПО РЕЧИ СОБЕСЕДНИКОВ

Двойникова А.А. (Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Университет ИТМО)

Научный руководитель – д.т.н., доцент Карпов А.А.

(Санкт-Петербургский Федеральный исследовательский центр Российской академии наук, Университет ИТМО)

В работе приводятся результаты экспериментальных исследований влияния размера окна спектрограммы на точность распознавания степени вовлеченности по аудиоданным собеседника. Исследования проводились на базе данных NoXi с применением сверточной нейронной сети.

Введение. В последнее время люди для коммуникации друг с другом активно используют информационные технологии. Это обусловлено удобством и комфортом общения между людьми, находящихся удаленно друг от друга. Для эффективной коммуникации собеседники используют как вербальные, так и невербальные сигналы общения. С помощью невербальных средств общения дикторы могут передавать различные состояния, такие как эмоции, настроение, интерес к диалогу, вовлеченность в него и пр. При живом общении понимание невербальных характеристик собеседника происходит довольно быстро и легко, т.к. собеседники могут видеть мимику, жесты, поведение коммуникантов. Однако при коммуникации посредством программного обеспечения, позволяющего в режиме реального времени видеть и слышать собеседника, анализ проявлений невербальных сигналов диктора становится затруднительным процессом, особенно если общение происходит между несколькими людьми одновременно. При проведении конференций, совещаний, лекций в дистанционном режиме основному диктору необходимо анализировать степень вовлеченности слушателей в разговор для того, чтобы понимать насколько интересна тема для собеседников, насколько сложен формат подачи материалов. Автоматическая система распознавания степени вовлеченности могла бы использоваться для решения данной проблемы. При виртуальном общении собеседники не всегда используют видеокамеру, в таких случаях анализ видеоданных для распознавания вовлеченности становится невозможным. В таких случаях анализ акустических характеристик речи собеседников может быть применим для автоматического распознавания вовлеченности участников разговора.

Основная часть. Для автоматического анализа вовлеченности по речи собеседников в работе использовалась база данных NoXi [Cafaro A. et al. The NoXi database: multimodal recordings of mediated novice-expert interactions //Proceedings of the 19th ACM International Conference on Multimodal Interaction. – 2017. – С. 350-359.]. Она содержит в себе аудио и видеозаписи участников разговоров на различные темы. Участники общались по парам на заранее выбранные темы по интересам, такие как спорт, политика, еда, технологии и т.д., причем один из участников выступал в роли эксперта в данной теме, а другой был новичком. Всего база данных содержит 84 диалога, между 87 участниками (61 мужчин и 26 женщин), возрастом 21–50 лет. Диалоги производились на различных языках, включая английский, французский, немецкий. Общая продолжительность базы данных составляет 25 ч 18 мин. NoXi содержит в себе непрерывную разметку уровня вовлеченности участника в диапазоне от 0 до 1. Для построения системы автоматического распознавания вовлеченности разметка группировалась по 4 классам: очень низкая вовлеченность [0-0,25), низкая вовлеченность [0,25-0,5), высокая вовлеченность [0,5-0,75) и очень высокая вовлеченность [0,75-1]. В каждом аудиофайле содержится речь одного диктора, однако дополнительно присутствует речь его собеседника, но с более низкой мощностью сигнала. Для того чтобы выделить речь необходимого диктора ко всем аудиозаписям применялся алгоритм VAD (англ. Voice Detection Activity –

обнаружение речевой активности), встроенный в библиотеку silero языка Python. Речь второго диктора принималась за шум и удалась из аудиозаписи. Затем аудиозаписи делились на окна длиной 40 мс (аннотаторы проводили разметку данных по аналогичным длительностям) с перекрытием 10 мс. Из каждого нарезанного аудиофрагмента извлекалась спектрограмма. Для экспериментальных исследований извлекались как узкополосные спектрограммы, так и широкополосные, отличающие разрешающей способностью полос. Для получения узкополосной спектрограммы кратковременное преобразование Фурье производилось со следующими параметрами: ширина спектральной полосы 10 Гц, окно Хэмминга и длина кадра исходного аудио 1 мс, а для широкополосной спектрограммы – ширина спектральной полосы 100 Гц, окно Хэмминга и длина кадра исходного аудио 20 мс. Затем данные спектрограммы подавались на вход сверточной нейронной сети (англ. Convolution Neural Network, CNN), архитектура которой состояла из 5 слоев сверток с нормализацией и пуллингом, 1 слоя дропаута и 1 полносвязного слоя.

Выводы. В работе исследуются зависимость точности распознавания степени вовлеченности диктора от ширины окна спектрограммы, извлекаемой из его акустических данных речи. Широкополосная спектрограмма хорошо отражает информацию о положение речевых формант, но плохо отображает гармоническую структуру речи. Узкополосная спектрограмма наоборот, представляет информацию о текстурных характеристиках звука. Использование узкополосных спектрограмм для анализа вовлеченности диктора по его речи показывает более высокую точность, чем при использовании широкополосных спектрограмм. Также разрабатывается автоматическая система распознавания степени вовлеченности речи диктора с применением сверточной нейронной сети.

Исследование выполнено при поддержке Совета по грантам Президента РФ (грант № НШ-17.2022.1.6)