Combinación de aprendizaje automático y análisis de escena auditiva computacional para separar el habla monoaural de dos hablantes

María Fernanda Toloza Buitrago
Ingeniería de sistemas
Deep learning

ENTRANDO EN CONTEXTO

La mayoría de los sistemas realizan agrupaciones basadas principalmente en el tono, como resultado, se limitan a segregar el habla sonora.

la técnica de aprendizaje automático tiene potencial para aplicarse a la separación de fuentes monoaurales.

En este artículo intentamos separar el habla sonora y no escrita simultáneamente mediante el uso del aprendizaje automático y alguna técnica relativa de CASA.

Enmascaramiento auditivo: dentro de una banda crítica, una señal más débil es enmascarada por una más fuerte

MÁSCARA BINARIA

La separación de voz podría tratarse como un proceso de agrupación

Clasificar el flujo de primer plano y el flujo de fondo correspondientes a la máscara 1 y 0, respectivamente

APRENDIZAJE AUTOMÁTICO + CASA

¿Cómo calcular las señales de enmascaramiento automáticamente a partir de una sola grabación de entrada?

1. Utilizar métodos de aprendizaje.
2. Descubrir estas regularidades a partir de una gran cantidad de datos.
3. Utilizar los modelos aprendidos para calcular las señales de enmascaramiento.
4. Realizar la separación.

CORPUS

Se compone de:
Conjuntos de capacitación.
Conjuntos de desarrollo.
Conjuntos prueba.

El conjunto de entrenamiento consta de 17000 frases (500 de cada uno de los 34 hablantes)

Tipo de datos:

oraciones con ruido en forma de habla

Los conjuntos de prueba contienen 600 frases en cada condición SNR.

los conjuntos de desarrollo tienen 300 rases en cada condición SNR.

Se manejaron archivos tipo wav a 25kHz

PROCEDIMIENTO

Preprocesamiento: extraer el cocleograma logarítmico del habla de entrada.

Entrenamiento: aprender el modelo correspondiente de cada hablante que se presenta en el corpus.

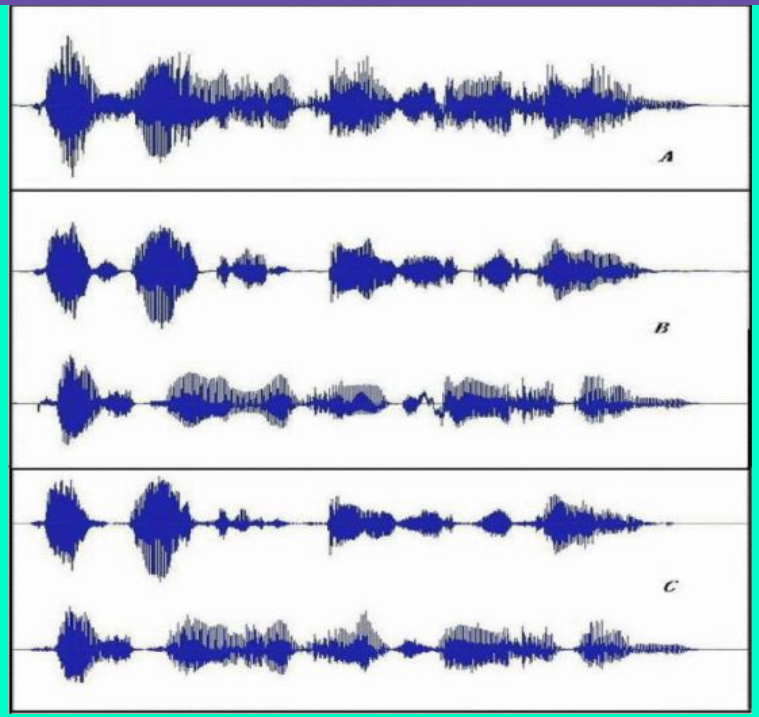
Prueba:al usar los libros de códigos de ambos hablantes y MAXVQ, podemos inferir las señales de enmascaramiento para el enunciado de prueba y realizar el procesamiento de resíntesis.

Preprocesamiento: extraer el cocleograma logarítmico del habla de entrada.

Entrenamiento: aprender el modelo correspondiente de cada hablante que se presenta en el corpus.

Prueba:al usar los libros de códigos de ambos hablantes y MAXVQ, podemos inferir las señales de enmascaramiento para el enunciado de prueba y realizar el procesamiento de resíntesis.

RESULTADOS



Forma de onda de los discursos (un caso de 0dB) (a) Forma de onda de la mezcla; (b) Forma de onda del objetivo original y la intrusión; (c) Forma de onda del objetivo separado y la intrusión

CONCLUSIÓN

En este artículo, se propone un método de separación de voz monoaural basado en aprendizaje automático y CASA. Mediante el uso de un paso de inferencia en un modelo factorial que se aprende de los datos de entrenamiento para proporcionar las señales de enmascaramiento para la resíntesis, se logra con éxito la separación de los discursos mixtos de dos hablantes. Los resultados de la evaluación del método propuesto sobre los datos de desafío de CASA muestran que el método propuesto podría separar muy bien el habla de dos hablantes y seguramente mejoró la SNR del habla de destino separada.

FUENTE:
https://ieeexplore-ieee-org.crai-ustadigital.usantotomas.edu.co/document/4368044

Published in: 2007 International Conference on Natural Language Processing and Knowledge Engineering