

Настройки распознавания речи с помощью движка Vosk

ОГЛАВЛЕНИЕ

1 (Общие сведения	3
2 E	Взаимодействие компонентов	4
3 A	Аппаратные требования	5
4 г	Порядок настройки	7



1 ОБЩИЕ СВЕДЕНИЯ

В LibraDLP имеется возможность использовать для распознавания речи движок **Vosk** – **ореп-source решение**, позволяющее **лучше масштабировать производительность** системы при распознавании аудиоданных.

Для использования **Vosk** не нужно приобретать лицензии – компонент не лицензируется. Это позволяет распознавать звуковые данные за **любой** период (в том числе архивные данные) из любого канала перехвата: *мессенджеры, звуковые файлы на внешних устройствах хранения (USB), звуковые файлы из папки (индексация локальных папок) и т.д.*

Интеграция инструмента в LibraDLP с поддержкой **различных языковых моделей** распознавания значительно **расширяет** возможности DLP-системы. При использовании Vosk количество потоков практически не влияет на использование оперативной памяти: для большой модели Vosk в среднем требуется 6-8 ГБ ОЗУ.

При использовании СРU один поток использует одно ядро. Например, на рабочей станции с 16 ядрами и 16 ГБ памяти можно максимально запустить 12 потоков, 3i Speech на этих же ресурсах будет работать в 1 поток. Также в сравнении с 3i Speech настройка движка Vosk значительно проще.

Для использования движка и просмотра результатов распознавания необходимо наличие следующих версий модулей LibraDLP:

- DataCenter 2.57.6.6 и выше.
- SearchServer 5.27.0.8 и выше.
- AAServer 2.3.0.20 и выше.
- AnalyticConsole 1.79.1.1 и выше.



2 ВЗАИМОДЕЙСТВИЕ КОМПОНЕНТОВ

Для взаимодействия компонентов системы должны быть открыты порты, представленные в Таблице 1.

Таблица 1 – Порты взаимодействия компонентов

Компонент	Протокол/Порт	Компонент	Протокол/Порт	
AAServer				
Взаимодействие с сере	вером баз данных	MS SQL		
Censen AAServer	TCP/*	Сервер баз данных MS	TCP/1433	
	UDP/*	SQL	UDP/1434	
Взаимодействие с сере	зером баз данных	PostgreSQL		
Сервер AAServer	TCP/*	Сервер баз данных PostgreSQL	TCP/5432	
Отправка и получение	данных с AAServe	r		
SearchServer	TCP/*	Сервер AAServer	TCP/9070	
AnalyticConsole				
Управление AAServer				
Клиент AAServer	TCP/*	Сервер AAServer	TCP/9070	
Клиент DataCenter	- ,		- ,	



3 АППАРАТНЫЕ ТРЕБОВАНИЯ

В зависимости от количества пользователей, аудиоданные которых необходимо распознавать, требуются следующие характеристики сервера для использования Vosk:

С малой моделью Vosk: до 10 пользователей CPU от 2.5 GHz (4 ядра) RAM от 8 GB HDD от 100 GB Microsoft Windows Server 2012 R2 и выше 0S Ubuntu 22.04 Astra Linux 1.7.3 до 20 пользователей CPU от 2.5 GHz (4 ядра) RAM от 8 GB HDD от 100 GB Microsoft Windows Server 2012 R2 и выше Ubuntu 22.04 OS Astra Linux 1.7.3 до 50 пользователей CPU от 2.5 GHz (8 ядер) RAM от 8 GB HDD от 100 GB Microsoft Windows Server 2012 R2 и выше Ubuntu 22.04 OS Astra Linux 1.7.3 до 100 пользователей CPU от 2.5 GHz (14 ядер) RAM от 8 GB HDD от 100 GB Microsoft Windows Server 2012 R2 и выше OS Ubuntu 22.04 Astra Linux 1.7.3

С большой моделью Vosk:

до 10 пользователей	
CPU	от 2.5 GHz (4 ядра)
RAM	от 16 GB
HDD	от 100 GB



	Microsoft Windows Server 2012 R2 и выше
OS	Ubuntu 22.04
	Astra Linux 1.7.3
до 20 пользователей	
CPU	от 2.5 GHz (4 ядра)
RAM	от 16 GB
HDD	от 100 GB
	Microsoft Windows Server 2012 R2 и выше
OS	Ubuntu 22.04
	Astra Linux 1.7.3
до 50 пользователей	
CPU	от 2.5 GHz (24 ядра)
RAM	от 16 GB
HDD	от 100 GB
	Microsoft Windows Server 2012 R2 и выше
OS	Ubuntu 22.04
	Astra Linux 1.7.3
до 100 пользователей	
CPU	от 2.5 GHz (48 ядра)
RAM	от 24 GB
HDD	от 100 GB
	Microsoft Windows Server 2012 R2 и выше
OS	Ubuntu 22.04
	Astra Linux 1.7.3



4 ПОРЯДОК НАСТРОЙКИ

В первую очередь в консоли <u>AAServer</u> на вкладке **Основные** проверьте наличие настроенного подключения к БД. Если подключение не настроено нажмите кнопку **Настройки подключения к БД** для создания новой базы данных, заполните параметры подключения к SQL и нажмите **Создать...** для создания БД настроек AAServer.

@	Additional Analysis Server Settings	? 🖬 🗕 🗖 🗙
localhost : 9070		Gi - 🗖 -
Остановить Перезагрузить Подключиться Обно сервер	ить	
Состояние служб		
Image: Weight of the state of th	Выключен Vosk Выключен Rassword Crack Выключен :: 0 / 0 △ Очередь / В обработке: 0 / 0 △ Очередь / В обработке: 0 / 0 △ Очередь / В обработке: 0 / 0 △	MailExport (split process merge) Выключен ередь / В обработке: 0 / 0 0 / 0 0 / 0 🛆
Параметры сервера и обработчиков		
Сервер / Обработчики Статус Адр	Параметры подключения к SQL Server X	ознавани Подбор Экспорт Время последней
ТW-LibraDLP Запущен 10.0	Настройки подключения SQL Server	паролеи почты активности
	задай те или сербера, тип и паралетры аутелтицикации, или од	
	Тип сервера:	
Основные ОСК Распознавание речи Подбор паролей Э	NSSQL 🔻	
Easa gallity	Имя сервера:	
	TW-LIBRADLP Считать из DC	
имя сервера	Использовать аутентификацию Windows	v
База данных	 Использовать внутреннюю аутентификацию SQL Server Имя пользователя: 	÷
Имя пользователя	sa ***********	*
Включить SSL-шифрование	Имя БД:	ф МБ
📴 Настройки подключения к БД	aaserver v Создать	💂 байт
🗖 Очистить базу данных 🕄 Удалить доку		
Папка с файлами С:\ProgramData\Librasoft\Librasoft AAServe	Проверить соединение ОК Отмена	цжеры: 0, НТТР (POST): 0, Устройство: 0, Ра
	* Значение "П" – и	
		Восстановить по умолчанию

Для извлечения текста из аудиофайлов с помощью движка **Vosk** выполните настройки:

1. В консоли <u>AAServer</u> на вкладке **Распознавание речи** включите опцию **Включить распознавание речи**.

Основные OCR Pacno	знавание речи Подбор паролей 3	кспорт почты	
Включить распознав	вание речи		
Vosk			
Таймаут задачи	5 🌲 минут		
Таймаут перезапуска	120 🌲 минут		
Кол-во потоков	4		
Модель	model-small-ru 🔻		
Параметры FFmpeg			
3i Speech			
Таймаут задачи	5 🌲 минут	Сервер	Файловый сервер
Кол-во потоков	4	Адрес 10.0.0.86	Адрес 10.0.1.27 👻
Модель	ru_broadcast_news_16000 🔻	Порт 50000 🗘	Порт 9080 🗘

2. Задайте параметры работы движка Vosk (см. таблицу ниже).



Таблица 2 – Параметры движка распознавания речи Vosk

Параметр	Описание				
Таймаут задачи	Максимальное время выполнения задачи, минут.				
Таймаут перезапуска	Время, по истечении которого движок будет перезапущен. Значение по умолчанию - 120 минут. Доступный диапазон: 30-240 минут. Если во время распознавания наступает время перезапуска по таймауту, сервер не перезапустится, пока не завершится распознавание. Если во время работы наблюдается чрезмерное потребление оперативной памяти, рекомендуется уменьшить значение тайм-аута.				
	Количество файлов, параллельно обрабатываемых каждым воркером. Рекомендованное число потоков равно 0,75 * кол-во ядер .				
Кол-во потоков	В связи с особенностями архитектуры Vosk для распределения нагрузки между ядрами для серверов на базе OC Windows существуют программные ограничения: если в системе более 30 ядер (логических процессоров), Vosk будет использовать только половину из них. Для серверов на базе OC Linux ограничений нет.				
	Набор файлов, определяющих уровень качества и язык распознавания. При установке на сервер устанавливается одна небольшая по объему модель model-small-ru , обеспечивающая высокую скорость обработки и базовый уровень качества.				
Модель	Из отдельных дистрибутивов можно установить дополнительные русские и английские модели , обеспечивающие более высокое качество распознавания. Установка дополнительных моделей осуществляется только на сервер! После этого на воркерах список моделей автоматически синхронизируется со списком на сервере. <u>Ссылка</u> на скачивание большой модели Vosk.				
	Также можно добавить в папку «C:\Program Files\Librasoft Librasoft AAServer\vosk\custom» любую доступную модель Vosk (можно <u>скачать</u> с официального сайта проекта), она будет использоваться при выборе из списка модели custom.				
Параметры FFmpeg	Дополнительные параметры для пакета FFmpeg для предварительной обработки распознаваемого звукового файла. По умолчанию применяются параметры afftdn=tn=enable и loudnorm .				

3. Перейдите в консоль <u>SearchServer</u> и включите использование движка Vosk. Для этого перейдите в меню Настройки → Экспертный режим → вкладка Опции → Сервисы извлечения текста → параметр Движок распознавания речи → Vosk. Нажмите Применить.



Unicode: Administrator console	-		x
	≷	-	00
Image: Construct C		¢r	
∠ Склфигурация сервера > Од 50,-сервер > Од детост лечатей > Од негост лечатей > Од негост лечатей > Од негост лечатей > Од неростры связиренного анализа			
Copercivit intervelenan Texcha Copercivit intervelenan Copercintervelenan Copercintervelenan Copercivit int			
Движнок распознавания речи vosk 5 Игнорируение классы изображений trash, temp 6 Классофикация Солтисти.1			
і Прикритет Низкий 5. Профиль ContentAI ТеxtExtraction_Accuracy			~

Появится уведомление о необходимости перезагрузки службы сервера для применения новых настроек. Нажмите **ОК**.

4. В этом же разделе ниже укажите адрес AAServer, с помощью которого будет осуществляться распознавание речи: в группе параметров Сервер расширенного анализа укажите в поле параметра Хост AAServer имя или IPадрес сервера с установленным AAServer, а в поле Порт AAServer – номер порта, на котором работает AAServer.

(онфиг	гура	ция сервера Переключитьс	я в обычный режи	MO		0 A B
Опция	A			Значение	2	^
	5	Отключение нормализации для символов	3			
	B	Переносы				
	Ś	Приоритет кодовых страниц		1251,866,20866	5,21866	
	5	Флаги поиска с опечатками				
	i.	сраничение поиска похожих		5 000		
	- i -	Опраничение размера массива для поиск	a	100 000		
🔺 🗋	Пар	аметры взаимодействия с DataCenter				
	В	Почтовые уведомления DataCenter		\checkmark		
	- i .	Язык почтовых уведомлений		Русский		
	5	Список получателей почтовых уведомле	ний			
-	- i -	Таймаут падключения к DataCenter		7 000		
_ ∧ D	Дет	ект печатей				
	В	Включить распознавание печатей		\checkmark		
	F	Качество распознавания		1		
	F	Порог классификации		40	-	
- Q	Cep	вер расширенного анализа				
	S	Xoct AAServer		localhost		
	i	Порт AAServer		9 070		
	S	Хост AAServer (только получение данны:	x)			
	i.	Порт AAServer (только получение данны	ix)	9 0 7 0		
		Таймаут полключения к AAServer		60 000		

- **5.** Появится уведомление о необходимости перезагрузки службы сервера для применения новых настроек. Нажмите **ОК**.
- Создайте индекс выбранного источника данных, если необходимого индекса нет (если индекс имеется, то перейдите к <u>шагу 7</u>). Для этого в SearchServer выберите из контекстного меню команду Создать индекс.



	🔚 🗮 🐻 🐻 🔜 Фильтр:	G D	R R G 2	
EC	Имя 🔺	Всего документо	в Проиндексирова	Дата об нов
📒 Индексы	🖻 🎰 FTP (7)			
🏇 Настройки	EC_FTP		- 108	06.04.2023.
	EC_writers_local_EC_FTP_AI~20211004		- 1	23.07.2024.
	6 FTP Создать индекс	Alt+C	-	23.07.2024.
				23.07.2024.
	Клонировать индекс		-	23.07.2024.
	— 👸 Inde: 🔜 Удалить индекс	Del		-
	Inde: 📴 Подключить индекс	Alt+M	-	-
	🖻 🜐 НТТР (🦰 Найти индексы	Alt+F		
		Alt. 11	- 14	17.02.2021.
	_NC	AIL+O	- 22 084	17.02.2021.
	. ЕС_Н Начать индексацию в порядке очер	еди Alt+S	- 66	17.02.2021.
	🗄 🚾 ЕС_ 🖳 Начать индексацию немедленно	Alt+I		23.07.2024.

Таблица 3 – Создание индекса

Действие

Шаг в мастере создания индекса

	S Создать индекс			x
	Создать новый индекс Укажите параметры нового индекса			4
Укажите Имя индекса, отметьте опцию Создать	Папка индекса: С:\ProgramData\Librasoft\Indexes\Index Иня индекса: <mark>Index</mark> Комментарии:			1
предварительной настройкой его параметров и выберите тип источника данных Другое.	Создать индекс с настройкани по умолчанию Создать индекс с предварительной настройкой его параметров (для опытных пользователей) Укажите данные, подлежащие индексации: Мессенджеры Почта Файловая окстена Удаленные файловые хранилища Ф Другое			
		Назад 🛛	Цалее	Отмена

	Создать индекс Источники данных Пожалуйста, выберите источники данных и укажите параметры подключения к базаи данных	×
Отметьте источник данных и нажмите кнопку Конфигурация. Выберите базу данных, которая будет индексироваться.	Конфигурация Иослючения Им Параметры © Облако Че указано> ••••••••••••••••••••••••••••••••••••	на



Далее в свойствах индекса переключитесь в экспертный режим и включите в группе Парсеры → Парсер аудио опцию Использовать AAServer.

При необходимости в параметре Список разрешений для анализа выберите типы файлов которые будут распознаваться.

🗿 Создать и	індекс		x
Свойства и Настройка	ндекса паранетров индекса		4
Конфигу	урация индекса <u>Переключиться в обычный режим</u> 🔅		¢ 🎦 🖺
Опция		Значение	^
* * * *	Парсер формата pdf Парсер формата wri Парсер формата wri Парсер формата eml Парсер формата eml Парсер формата msg Парсер формата msg Парсер формата hlp Парсер удио і Мининальный размер индексируемого файла (байты) і максимальный размер индексируемого файла (байты)	-1 -1	
	В Использовать AAServer 5 Список расширений для анализа	way:mp3:opus:flac	
•	Парсер видео	nar inipolopios	
•	Парсер векторной графики		
+	Парсер текста		*
		Назад	Далее Отмена

Нажмите Далее .	6 Создать индекс	x
	Типы файлов Выберите типы файлов для индекса	4
	Выберите типы файлов для индекса Добавить Удалить Выбрать все Очистить	
	✓ Исключать	
	B Old text formats (2) Hasag Aanee OTHE	на

	😉 Создать индекс			x
При необходимости добавьте расписание	Расписание Настройка расписания индексации Автообновление индекса будет запускаться по настроенному расписанию Добавить Удалить Редактировать			ł
индексации для	Название расписания	Параметры расписания	Следующий старт	
данного индекса и убедитесь, что установлен флажок Запустить индексацию после создания индекса.	Расписание1	Ежедневно	23.07.2024 12:45:48	
Нажмите Завершить .	✓ Запустить индексацию после создания инде	екса	Назад Завершить Оти	ена



7. В свойствах выбранного/созданного индекса на вкладке Опции в группе Парсер аудио включите опцию Использовать AAServer.

6 Свойс	тва инде	ca - Index								x
Общие	Опции	Типы файлов	Расписание	Мониторинг						
Конф	игурац ия	ия индекса	<mark>Перекл</mark>	ючиться в обы	иный режим 🗔	Значение			¢@B ^	
	▶ 🗋 Па	арсер web-архиво	ов (mht)							
	🕨 🗋 Па	арсер формата ht	ml							
	🕨 🗋 Па	арсер формата ро	If							
	🕨 🗋 Па	арсер формата wr	i							
	🕨 🗋 Па	арсер офисных до	окументов							
	🕨 🗋 Па	арсер формата еп	nl							
	🕨 🗋 Па	арсер почтовых а	рхивов (pst)							
		арсер формата ms	sg							
	Парсер формата hlp									
	▲ [] <u>⊓</u> a	арсер аудио								
	i	Минимальный р	размер индекси	руемого файла	(байты)	-1				
	- E	Максимальный	размер индекс	ируемого файла	а (байты)	-1				
	Ľ	Использоват	ь AAServer				-			
		Список расшир	ений для анали	138		wav;mp3;opus;	пас			
		арсер видео								
	P	арсер векторной і	графики						*	
								ОК	Отмена	

На этом необходимые настройки для распознавания речи с использованием движка Vosk в консоли SearchServer завершены.

- **8.** После этого требуется подключить (смонтировать) созданный индекс в консоли <u>EndpointController</u>. Для этого перейдите в раздел SearchServer.
- **9.** Выделите сервер, на котором установлен компонент SearchServer, и щелкните кнопку **Смонтировать индексы**. В качестве альтернативного варианта можно выбрать команду **Смонтировать индексы** из контекстного меню.

Индексы EndpointC Конфигурация индексов Е	Controller EndpointController	
Search Server 🔺		
Перехват на агентах		
	Создать индексы	
E Cloud_AI	Смонтировать индексы	
🗄 📃 Device_Al	Демонтировать индекс	1
E Device_A	Удалить индекс	
	Остановить сервер	
	Удалить сервер	
HTTP_AI^	Перенести активные индексы	
нформация о Search Serv	Начать обновление	
Search Server	Очистить индекс	
Версия протокола:	Добавить источник данных	-
Состояние сервера:	Удалить источник данных	
Индексов:	26	1
Создать индексы	монтировать индексы Остановить серве;	Удалить сервер

Отметьте флажком созданный ранее индекс. Чтобы перехват сохранялся в выбранном индексе, установите флажок Сделать индексы активными.



🚯 Подключение существующих индексов	x				
Подключить индексы в программу Выберите индексы для подключения					
Search Server: EC:8953 🗸					
🔚 🗮 🛛 🐺 Перехват на агентах 🗸 Поиск:					
Список доступных индексов Описание					
	~				
Cloud					
Cloud_AI					
Cloud_AI~20210713_ver56404					
EC_Cloud_SharePoint Imported from SearchServer 4.xS					
Index					
DeviceSnifferDS;DeviceSnifferDS;DeviceSnifferDS (4)					
🖭 🗌 🔞 FileAuditor (4)					
	\mathbf{r}				
Сделать индексы активными Добавить Отме	на				

Смонтированные индексы будут отображены в окне просмотра.

