UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA DE SISTEMAS E INDUSTRIAL
MAESTRÍA EN INGENIERÍA DE SISTEMAS

DEIBY ALEXANDER FANDIÑO RODRÍGUEZ
COD 256349

BIBLIOGRAFÍA ANOTADA

Article
José A. Brito, W.J.R.
Identificación de Señales Verbales en el Espacio de Fase Reconstruido Universidad de Los Andes, Postgrado en Computación,, ***1999***

Abstract: In this paper we describe the use of Multilayer Perceptron Array for learning and classifying speech signals, using characteristic vectors of reconstructed dynamics. First, we consider the phonatory system as a black-box, where the only available data is its output: the speech signal. Theoretically, if reconstruction of system dynamics is properly made, geometric structures or attractors outlined in the space are topologically equivalent to original, and inaccessible, structures. This is a way of accessing underlying dynamics, and is the starting point for two kinds of experiments: classification of vowels and digits, with Venezuelan Spanish voices. Results verify positively that characteristics vectors extracted from underlying dynamics hold discriminative power for distinguishing between classes of speech signals. Besides, neural networks are able to generalize using this kind of data. Keywords: Speech signals classification, reconstructed dynamics, pattern recognition, non linear dynamics, neural nets, SpeechDat. Resumen Este artículo se describe el uso de arreglos de redes neuronales de retropropagación para el aprendizaje y clasificación de señales verbales, usando vectores de características de la dinámica reconstruida. Primero, se considera el sistema fonatorio como una caja negra, donde la única data disponible es la salida: la señal verbal. Teóricamente, si la reconstrucción de la dinámica del sistema es correcta, las estructuras geométricas o atractores del espacio son topologicamente equivalentes a las estructuras originales inaccesibles. Esta es una forma de acceder a la dinámica subyacente, y es el punto de partida para dos tipos de experimentos: clasificación de vocales y dígitos, con voces en español venezolano. Los resultados verifican positivamente que los vectores de características extraídos de la dinámica subyacente tiene poder discriminatorio para distinguir entre clases de señales verbales. Además, las redes neuronales son capaces de generalizar usando este tipo de datos. Palabras claves: Clasificación de

señales verbales, espacio de fases reconstruido, reconocimiento de patrones, dinámica no lineal, redes neuronales, SpeechDat.

Resumen:
Este artículo se describe el uso de arreglos de redes neuronales de retropropagación para el aprendizaje y clasificación de señales verbales, usando vectores de características de la dinámica reconstruida. Primero, se considera el sistema fonatorio como una caja negra, donde la única data disponible es la salida: la señal verbal.

Teóricamente, si la reconstrucción de la dinámica del sistema es correcta, las estructuras geométricas o atractores del espacio son topologicamente equivalentes a las estructuras originales inaccesibles. Esta es una forma de acceder a la dinámica subyacente, y es el punto de partida para dos tipos de experimentos: clasificación de vocales y dígitos, con voces en español venezolano. Los resultados verifican positivamente que los vectores de características
Extraídos de la dinámica subyacente tiene poder discriminatorio para distinguir entre clases de señales verbales.

El artículo es de mucha relevancia con el tema debido a que profundiza sobre el tema del tratamiento de señales con redes neuronales, además los resultados son claros.

Article
Ovidiu Grigore, I.G.
Neuro-fuzzy Models for Speech Pattern Recognition in Romanian Language Polytechnic University of Bucharest, Dept. Electronic and Telecommunications, ***1998***

Abstract: in this paper are presented results obtained in a vowel recognition task applying fuzzy neural networks. The vowels, uttered from 10 speakers each in 1000 different contexts are recognized using as features the first three formant frequencies. The results obtained in this case show that fuzzyfication process improves the recognition rate of the classical variants. The paper is organized in 6 chapters: after an introduction (1), the feature extraction (2) is presented. The two fuzzy neural networks: the fuzzy multilayer perceptron (2) and the fuzzy Kohonen map (3) used for recognition are given. Conclusions about the obtained results with future plans (4) and a reference list (5) close the paper.

Resumen:

El artículo resume los resultados de las investigaciones sobre la implementación de las redes neuronales junto con la lógica difusa para el reconocimiento de palabras.

Las pruebas se realizaron con un corpus de datos recopilados de 10 personas en más de mil contextos diferentes, la red neuronal que se entreno fue una red de Kohonen de 4 capas y los resultados fueron comparados con una red MLP.

Como resultado se vio que la red que tuvo mas acierto fue en la que se implemento la capa de lógica difusa, ya que el porcentaje de error disminuyo notablemente, y se llego a la conclusión de que se puede crear un sistema dinámico para el reconocimiento de palabras.

Inproceedings
Ha-Jin Yu, Y.O.
A NEURAL NETWORK USING ACOUSTIC SUB-WORD UNITS FOR CONTINUOUS SPEECH RECOGNITION
1999

Abstract: A subword-based neural network model for continuous speech recognition is proposed. The system consists of three modules, and each module is composed of simple neural net- works. The speech input is segmented into non-uniform units by the network in the rst module. Non-uniform unit can model phoneme variations which spread for several phonemes and between words. The second module recognizes seg- mented units. The unit has stationary and transition parts, and the network is divided according to the two parts. The last module spots words by modeling temporal representation. The results of speaker independent word spotting of 520 words are described.

Resumen:

El articulo describe el procedimiento que se utiliza para el reconocimiento de palabras en un discurso continuo, lo cual presenta una dificultad y es la influencia que tienen un fonema al pronunciar el siguiente, el proceso de reconocimiento se dividió en dos módulos , el primero consiste en segmentar todas las palabras en fonemas independientes, para luego entrenar la red y así simplificar el proceso de selección e identificación, el segundo modulo realiza el proceso de entrenamiento de la red e identificación de los fonemas para así identificar las palabras completas.

El trabajo presenta análisis de resultados importante ya que por medio de las pruebas realizadas concluyen que al segmentar el conjunto de palabras el margen de error de reconocimiento disminuye considerablemente, además este sistema se esta comparando con otros en la actualidad.

Inproceedings
Nicolas Pican, D.F.
HMMs and OWE Neural Network for Continuous Speech Recognition
2001

Abstract: The phonetic context has a large effect on stop consonants in a continuous speech signal [1]. Therefore recognition systems that model allophones using context-dependent Hidden Markov Models have been implemented [3]. HMMs have a great ability for the segmentation in the temporal domain [4][6] but have some difficulties in the recognition because the MLE training (Maximum Likelihood Estimation) is not discriminant, whereas the discrimination is one of the abilities of the Artificial Neural Networks models. In the last three years we have developed a new ANN model named OWE (Orthogonal Weight Estimator)[9][10]. The principle of the OWE is a ANN that classifies an input pattern according to contextual environment. This new ANN architecture tackles the problem of context dependent behaviour training. Roughly, the principle is based on main MLP (Multilayered Perceptron) in which each synaptic weight connection value is estimated by another MLP (an OWE) with respect to context representation. In this paper, we present a hierarchical system for phoneme recognition: first the system segments the input signal using 48 context independent HMMs. Then the stop consonant are reordered by a OWE ANN. Experiments on TIMIT show 78 % of correct recognition rate on the 6 stop consonants (/p, t, k, b, d, g).

Resumen:
El articulo presenta la investigación del proceso de reconocimiento de palabras en un discurso continuo, explica que el estándar en este tipo de procesos de reconocimiento es utilizando modelos ocultos de Markov , en los cuales las palabras son independientes del contexto.

Explican por que la red que ellos entrenan (DEBA) simplifica aun mas el problemas de selección y como el error al reconocer disminuye considerablemente comparado con otras técnicas. Ellos presentan un sistema de reconocimiento de fonema jerárquico basado en modelos HMM y DEBE. La clasificación se realiza por segmentación temporal.

Concluyen que los resultados más buenos se han obtenido cuando el entrenamiento es hecho con segmentación de HMM que no introduce ningún prejuicio entre probar y entrenar las condiciones. Se demuestra también que las grandes capacidades de la arquitectura DEBA en el contexto - clasificación dependiente que es una tarea difícil en el reconocimiento del discurso.

Inproceedings
Jordi Adell, A.B.
Análisis de la Segmentación Automática de Fonemas para la Síntesis de Voz. ***2001***

Abstract: En este artículo se presentan dos nuevos sistemas para la segmentación de voz en fonemas. Uno basado en un clustering acústico previo a un alineado por programación dinámica y el segundo basado en una corrección especifica de las fronteras mediante un árbol de regresión. Se discute el uso de medidas objetivas o preceptúales para la evaluación de estos sistemas. Los sistemas presentados claramente mejoran los resultados del

sistema de partida basado en HMM y obtienen resultados similares a la concordancia entre diferentes segmentaciones manuales. Se muestra como las características fonéticas pueden ser utilizadas satisfactoriamente, junto con los HMM, para la detección de las fronteras. Finalmente, se enfatiza la necesidad de utilizar tests preceptúales para evaluar la segmentación de las bases de datos para síntesis de voz.

Resumen:

En el artículo se revisan las técnicas existentes destinadas a la segmentación de voz aplicadas a la síntesis. También se discuten los métodos posibles para la evaluación de estas técnicas y sobre la conveniencia de realizar o no evaluaciones objetivas y/o preceptúales. El hecho de aplicar una serie de sistemas a una misma base de datos y bajo las mismas condiciones concluye que se puede comparar de forma adecuada los diversos sistemas.

Se presentan dos nuevos métodos para la segmentación que han mejorado los resultados del sistema de partida basado en HMM. Los resultados muestran que es posible conseguir resultados similares a las discrepancias entre segmentaciones manuales, simplemente refinando las fronteras en base a las características fonéticas usando un árbol de decisión.

Inproceedings
Imael Cortázar Múgica, M.Á.R.C.
Últimos desarrollos en tecnologías de voz y del lenguaje
2002

Abstract: *El objetivo de este artículo es presentar los últimos desarrollos realizados en Telefónica I+D de las tecnologías de voz y del lenguaje (reconocimiento de voz, verificación del locutor, procesamiento del lenguaje natural, agentes inteligentes aplicados a la gestión del diálogo y conversión texto-voz), así como las posibles aplicaciones de estas tecnologías a los servicios reales ofrecidos a través de la línea telefónica. Para ello, se describen estos desarrollos y se indica que nuevos servicios, o que mejoras en los actuales, tienen la posibilidad de ser ofertados.*

Resumen:

El articulo presenta una recopilación de varios estudios realizados por el centro de investigación I+D, en el área de las tecnologías de voz y lenguajes. Presentan un sistema que permite el diálogo con el usuario en tiempo real a través de terminales telefónicos fijos o móviles. Los resultados obtenidos han permitido confirmar las ventajas del uso de la tecnología de agentes.

Article
Eduardo Clemente, A.V.

Entrenamiento y Evaluación de reconocedores de Voz de Propósito General basados en Redes Neuronales feedforward y Modelos Ocultos de Markov TLATOA-CENTIA, ***1999*** *, 15*

Abstract: Este artículo presenta los resultados de un trabajo de tesis que consistió en el desarrollo y comparación de dos nuevos reconocedores de propósito general para el español Mexicano. Uno de estos reconocedores está basado en la metodología de redes neuronales y el otro en modelos ocultos de Markov. El interés en la comparación de estas metodologías se debe a que sólo se habían comparado reconocedores de propósito específico (dígitos) [2, 8], de ahí que el desarrollo de un reconocedor de propósito general permitirá realizar cualquier tipo de aplicación para el español hablado en México.

Resumen:

El artículo resume la investigación realizada, sobre la implementación de un reconocedor fonético para dígitos en español, este proyecto fue implementado usando el Toolkit, el cual es un paquete de aplicaciones desarrollado específicamente para sistemas de reconocimiento de voz.

Se explica el proceso para la clasificación de los fonemas, el cual reduce el problema y permite una mayor probabilidad que el sistema lo interprete correctamente, después de este proceso se sigue con el entrenamiento de la red neuronal de arquitectura feed forwad de tres niveles que emplea el algoritmo back propagation. El proceso consistió en dividir la señal (fonema) en frames, se calculan los vectores de características a partir del espectro de la señal (voz), se clasifican teniendo en cuenta características como energía y anchos de frecuencia, y después estos vectores se introducen a la red neuronal. Después de entrenar la red se paso a la etapa de desarrollo, en la cual se selecciona la red con mejor desempeño, lo cual se realiza evaluando el nivel de error en el proceso de reconocimiento.

Inproceedings
Merlo, G.F.
Reconocimiento De La Voz Mediante Una Red Neuronal De Kohonen ***1997***

Abstract: El reconocimiento de la voz mediante diversas técnicas tales como cadenas ocultas de Markov y Redes Neuronales es tema de investigación constante, obteniendo resultados de distinta performance según el método elegido. En el presente trabajo se exhiben los resultados de una experiencia en reconocimiento de voz de un individuo, tomando como patrones a ser reconocidos las cifras decimales (0- 9), y utilizando como método una red neuronal de Kohonen. Luego de una fase de entrenamiento y sintonización, produce con solo cien neuronas un aceptable resultado de reconocimiento (65%).

Resumen:

El artículo describe básicamente el proceso de entrenamiento una red neuronal de kohonen para el reconocimiento de la voz, consta de tres partes bien diferenciadas:
Grabación de los archivos sonoros y transformación de los mismos en un conjunto de datos o patrones que sean entendibles por parte de la red. Esta tarea de mapeo de datos es realizada por un bloque codificador de entrada.
B) Generación y entrenamiento de la red con los patrones. Este proceso constituirá la red neuronal propiamente dicha.
C) Prueba de la red neuronal obtenida luego del aprendizaje, mediante el reconocimiento de nuevas ocurrencias de los archivos de voz.

El trabajo muestra un aceptable comportamiento de la red neuronal de Kohonen frente a la tarea de reconocimiento de un patrón tan estocástico como lo es la voz. Aunque su tasa de acierto dista aún mucho de las obtenidas mediante otros métodos, existen muchas variables en la red neuronal que pueden ser ajustadas para lograr mejores comportamientos de la misma

Article
N.Munive, A.v.
Entrenamiento de un reconocedor fonético de dígitos para el español de México usando CSLU TOOLKIT
Computación y sistemas, ***1999*** *, 3*

Abstract: Este trabajo presenta el diseño de un reconocedor fonético de dígitos para el español hablado en México, implementado usando el toolkit.
Resumen:

El artículo resume la investigación realizada, sobre la implementación de un reconocedor fonético para dígitos en español, este proyecto fue implementado usando el Toolkit, el cual es un paquete de aplicaciones desarrollado específicamente para sistemas de reconocimiento de voz. La investigación se centra en 3 partes: Creación del corpus Desarrollo del clasificador Evaluación de resultados El corpus comprende toda la base de conocimiento del programa, esto incluye la colección de los fonemas y las palabras del grupo de muestra tomado, que en este caso fue de 50 personas, 25 mujeres y 25 hombres, el resultado de este proceso fue la colección de 1933 archivos de voz. Desarrollo del clasificador En esta fase se explica el proceso para la clasificación de los fonemas, el cual reduce el problema y permite una mayor probabilidad que el sistema lo interprete correctamente, después de este proceso se sigue con el entrenamiento de la red neuronal de arquitectura feed forwad de tres niveles que emplea el algoritmo back propagation. La red neuronal esta constituida por 130 nodos de entrada , 200 nodos ocultos y un nodo de salida el cual es la categoría fonética que se desea reconocer. El proceso consistió en dividir la señal (fonema) en frames, se calculan los vectores de características a partir del espectro de la señal (voz), se clasifican teniendo en cuenta características como energía y anchos de frecuencia, y después estos vectores se introducen a la red neuronal. Después de entrenar la red se paso a la etapa de desarrollo,

en la cual se selecciona la red con mejor desempeño, lo cual se realiza evaluando el nivel de error en el proceso de reconocimiento. Se desarrollaron 3 clasificadores: Clasificador independiente de contexto Clasificador dependiente de contexto Clasificador dependiente del contexto agrupando los fonemas en clases generales Para esto se implementaron las pruebas con un corpus grabado por teléfono.

Se observa un proceso estructurado y claro, lo cual facilita el entendimiento del artículo.

No presenta un marco teórico, y omite definiciones, las cuales da por hecho que el lector tiene conocimiento.

Los resultados descritos son claros, y concluyentes, sin embargo la forma en que son presentados es muy simple, se podrían haber utilizado recursos como graficas para las comparaciones.

Se ignoran datos de los tiempos de procesamiento de las redes, así como el número de pruebas que se realizaron, si se realizaron más de una.

Conference
PETRUSHIN, V.A.
Emotion In Speech:Recognition And Application To Call Centers ***1999***

Abstract: The paper describes two experimental studies on vocal emotion expression and recognition. The first study deals with a corpus of 700 short utterances expressing five emotions: happiness, anger, sadness, fear, and normal (unemotional) state, which were portrayed by thirty non-professional actors. After evaluation a part of this corpus was used for extracting features and training backpropagation neural network models. Some statistics of the pitch, the first and second formants, energy and the speaking rate were selected as relevant features using feature selection techniques. Several neural network recognizers and ensembles of recognizers were created. The recognizers have demonstrated the following accuracy: normal state - 60-75%, happiness – 60- 70%, anger – 70-80%, sadness – 70-85%, and fear – 35-55%. The total average accuracy is about 70%. The second study uses a corpus of 56 telephone messages of varying length (from 15 to 90 seconds) expressing mostly normal and angry emotions that were recorded by eighteen non-professional actors. These utterances were used for creating recognizers using the methodology developed in the first study. The recognizers are able to distinguish between two states: “agitation” which includes anger, happiness and fear, and “calm” which includes normal state and sadness with the average accuracy 77%. An ensemble of such recognizers was used as a part of a decision support system for prioritizing voice messages and assigning a proper agent to response the message. The architecture of the system is presented and discussed.

Inproceedings
Taylor, J.F.K.R.S.K.P.
AN AUTOMATIC SPEECH RECOGNITION SYSTEM USING NEURAL NETWORKS AND LINEAR DYNAMIC MODELS TO RECOVER AND MODEL ARTICULATORY TRACES.

Abstract: We describe a speech recognition system which uses articulatory parameters as basic features and phone-dependent linear dynamic models. The system first estimates articulatory trajectories from the speech signal. Estimations of x and y coordinates of 7 actual articulator positions in the midsagittal plane are produced every 2 milliseconds by a recurrent neural network, trained on real articulatory data. The output of this network is then passed to a set of linear dynamic models, which perform phone recognition.
El articulo presenta el estudia que se realizo para la el reconocimiento de patrones de voz para establecer el estado de animo de la persona que esta hablando. El estudio comprende de varias fases, una en la que se recopila toda la información para entrenar la red (corpus), otra en la que se entrena la red neuronal y otra en la que se analizan los resultados obtenidos.

Resumen:

El artículo describe el procedimiento para el reconocimiento de sonidos polifónicos y monofónicos, en un contexto dependiente del contexto en este caso las pruebas y el corpus de datos se realizaron por sonidos generados por la línea telefónica.

Se explica el proceso de descomposición de las señales en vectores, esto se realiza utilizando la teoría de tratamiento de señales, una vez que se ha descompuesto la señal esta se clasifica en vectores para facilitar el proceso de reconocimiento.

El artículo expone la importancia de los resultados, ya que esta técnica puede ser utilizada para el reconocimiento de sonidos más complejos e incluso explican la viabilidad de utilizarla para el proceso de reconocimiento del habla humana en cualquier contexto.

Mastersthesis
Ahuactzin, I.K.N.A.A.
*Aplicación de Tecnología de Voz en la Enseñanza del Español Universidad de las Américas- Puebla., **2001***

Abstract: Este artículo presenta dos herramientas desarrolladas con tecnología de voz para la enseñanza del español hablado en México. Estas son un diccionario Inglés-Español accesado por medio de la voz, y un nuevo método para la verificación de la pronunciación correcta de palabras o frases. Esta

segunda herramienta es especialmente útil en los sistemas para la enseñanza de un lenguaje por medio de la computadora (CALL - Computer Aided Language Learning). Este método aprovecha las técnicas de reconocimiento de voz y las herramientas del CSLU Toolkit (del Center for Spoken Language Understanding, Oregon Graduate Institute) para reconocer la secuencia de sonidos emitidos por el usuario y marcar las partes mal pronunciadas. Cada frase o palabra pronunciada por el usuario puede evaluarse fonema por fonema detectando los errores en que incurrió el locutor. Para ello es necesario entrenar un sistema de reconocimiento de voz (una red neuronal en este caso) con los fonemas del lenguaje objetivo además de incluir los fonemas del idioma nativo del locutor. Esto se debe a que los errores de pronunciación de alguien que esta aprendiendo un nuevo idioma son causados por las costumbres de pronunciación del locutor en su lengua materna. A diferencia de otros sistemas existentes, con este método se puede proveer una retroalimentación explícita y clara al usuario, el cual podrá entonces estudiar e intentar pronunciar las frases que contengan esas partículas que requiere practicar.

nproceedings
Sherif Yacoub, S.S.
Recognition of Emotions in Interactive Voice Response Systems ***2003***

Abstract: This paper reports emotion recognition results from speech signals, with particular focus on extracting emotion features from the short utterances typical of Interactive Voice Response (IVR) applications. We focus on distinguishing anger versus neutral speech, which is salient to call center applications. We report on classification of other types of emotions such as sadness, boredom, happy, and cold anger. We compare results from using neural networks, Support Vector Machines (SVM), K-Nearest Neighbors, and decision trees. We use a database from the Linguistic Data Consortium at University of Pennsylvania, which is recorded by 8 actors expressing 15 emotions. Results indicate that hot anger and neutral utterances can be distinguished with over 90% accuracy. We show results from recognizing other emotions. We also illustrate which emotions can be clustered together using the selected prosodic features.

Resumen:

El articulo presenta el estudio que se realizo para la el reconocimiento de patrones de voz para establecer el estado de animo de la persona que esta hablando. El estudio comprende de varias fases, una en la que se recopila toda la información para entrenar la red (corpus), otra en la que se entrena la red neuronal y otra en la que se analizan los resultados obtenidos.

Se tratan de establecer tipos de emociones como la tristeza, fastidio, felicidad y enojo, para poder realizar esta identificación, las palabras reconocidas se dividen en frames, así se pueden clasificar sus fonemas en vectores los cuales

se pueden diferenciar por su señal, estos vectores son los que ingresan a la red neuronal para realizar la clasificación.

Article
J.L. Gauvain, L.L.
Conversational Telephone Speech Recognition
IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), ***2003*** *, 212-215*

Abstract: This paper describes the development of a speech recognition system for the processing of telephone conversations, starting with a state-of-the-art broadcast news transcription system. We identify major changes and improvements in acoustic and language modeling, as well as decoding, which are required to achieve state-of-theart performance on conversational speech. Some major changes on the acoustic side include the use of speaker normalization (VTLN), the need to cope with channel variability, and the need for effi- cient speaker adaptation and better pronunciation modeling. On the linguistic side the primary challenge is to cope with the limited amount of language model training data. To address this issue we make use of a data selection technique, and a smoothing technique based on a neural network language model. At the decoding level lattice rescoring and minimum word error decoding are applied. On the development data, the improvements yield an overall word error rate of 24.9% whereas the original BN transcription system had a word error rate of about 50% on the same data.

Resumen:

El artículo presenta la investigación del proceso de reconocimiento de palabras en una conversación continua, en un ambiente dependiente (telefónico).

Para el desarrollo del identificador se creo un diccionario con las pronunciaciones de las palabras las cuales se utilizarían en la conversación, estas palabras de dividieron en clase fonéticas, después de esto se hace los mismo con las palabras de los interlocutores y se comparan con las del diccionario.

Inproceedings
Yukyz, D. & Flanagany, J.
TELEPHONE SPEECH RECOGNITION USING NEURAL NETWORKS AND HIDDEN MARKOV MODELS
2000

Abstract: The performance of well -trained speech regognizer, using high quiality full bandwith speech dara is usually degraded when used in real world environments.

En el articulo se exhiben los resultados de una experiencia en reconocimiento de voz de un individuo, tomando como patrones a ser reconocidos las cifras decimales (0-9), y utilizando como método una red neuronal de Kohonen. Luego de una fase de entrenamiento y sintonización, produce con solo cien neuronas un
Aceptable resultado de reconocimiento (65%).

El trabajo muestra un aceptable comportamiento de la red neuronal de Kohonen frente a la tarea de reconocimiento de un patrón tan estocástico como lo es la voz. Aunque su tasa de acierto dista aún mucho de las obtenidas mediante otros métodos, existen muchas variables en la red neuronal que pueden ser ajustadas para lograr mejores comportamientos de la misma; entre ellas, se pueden considerar:
Número de neuronas de la red.
Dimensión de los patrones y de dichas neuronas.
Cantidad de locutores y de archivos sonoros utilizados en el entrenamiento.
Método de codificación o mapeo elegido para la aplicación
Valores de los coeficientes de aprendizaje y sintonización, y forma de decaimiento en el
Tiempo.

Conference
Ries, K.
HMM AND NEURAL NETWORK BASED SPEECH ACT DETECTION ***1999***

Abstract: We present an incremental lattice generation approach to speech act detection for spontaneous and overlapping speech in telephone conversations (CallHome Spanish). At each stage of the process it is therefore possible to use different models after the initial HMM models have generated a reasonable set of hypothesis. These lattices can be processed further by more complex models. This study shows how neural networks can be used very effectively in the classification of speech acts. We find that speech acts can be classified better using the neural net based approach than using the more classical ngram backoff model approach. The best resulting neural network operates only on unigrams and the integration of the ngram backoff model as a prior to the model reduces the performance of the model. The neural network can therefore more likely be robust against errors from an LVCSR system and can potentially be trained from a smaller database.

Conference
Huckvale, M.
Phonetic characterisation and lexical access in non-Segmental speech recognition

1995

Abstract: An isolated-word speech recognitionsystem, built without the use of linearsegments for acoustic modelling or lexicalaccess, is justified, described and demonstrated. The system comprises phonetic feature analysis operating on four independent tiers, parallel phonotactic parsing, and lexical access based on a neural-network inspired lexicon structure. Performance is however still inferior to a baseline segmental system.

Inproceedings
Lin Zhong, Y.S. & Liu, R.
A Dynamic Neural Network for Syllable Recognition
1994

Abstract: A dynamic neural network architecture based on the Time-Delay Neural Network and the Convolutional Neural Network is originated. The dynamic network achieves much better performance than those of MLP and TDNN when dealing with syllable recognition. Such performance is also comparable to that of the more popular HMM method.

El articulo presenta el estudio realizado para desarrollar un reconocedor de silabas, basado en el metodo HMM, el cual se basa en la descomposición del fonemas en frames utilizando la teoría de tratamiento de señales la cual es muy amplia y ya lleva muchos años en practica.

El proceso consistió en dividir la señal (fonema) en frames, se calculan los vectores de características a partir del espectro de la señal (voz), se clasifican teniendo en cuenta características como energía y anchos de frecuencia, y después estos vectores se introducen a la red neuronal.

En los resultados se expone la efectividad del metodo ya que reconoció 408 silabas para un porcentaje de efectividad del 97% el cual es muy favorable para la complejidad del proceso.

Inproceedings
Strachan, S.K.T.S.S.I.P.T.A.
SPEECH RECOGNITION VIA PHONETICALLY FEATURED SYLLABLES
2001

Abstract: We describe a speech recogniser which uses a speech productionmotivated phonetic-feature description of speech. We argue that this is a natural way to describe the speech signal and offers an efficient intermediate parameterisation for use in speech recognition. We also propose to model this description at the syllable rather than phone level. The ultimate goal of this work is to generate syllable models whose parameters explicitly describe the trajectories of the phonetic features of the syllable. We hope to

move away from Hidden Markov Models (HMMs) of context-dependent phone units. As a step towards this, we present a preliminary system which consists of two parts: recognition of the phonetic features from the speech signal using a neural network; and decoding of the feature-based description into phonemes using HMMs.

Resumen:

El articulo presenta la investigación del proceso con el cual se realiza el reconocimiento de silabas, explica que el estándar en este tipo de procesos de reconocimiento es utilizando modelos ocultos de Markov, en los cuales las palabras son independientes del contexto.

Explican por que la red que ellos entrenan (HMM) simplifica aun mas el problemas de selección y como el error al reconocer disminuye considerablemente comparado con otras técnicas. Ellos presentan un sistema de reconocimiento de fonema jerárquico basado en modelos HMM y modelos ocultos de Markov. La clasificación se realiza por segmentación temporal.

Concluyen que los resultados más buenos se han obtenido cuando el entrenamiento es hecho con segmentación de HMM que no introduce ningún prejuicio entre probar y entrenar las condiciones. Se demuestra también que las grandes capacidades de la arquitectura HMM en el contexto - clasificación dependiente que es una tarea difícil en el reconocimiento del discurso.

Inproceedings
Joaquim Llisterri, C.C.
La conversión de texto en habla: aspectos lingüísticos ***2000***

Abstract: Un sistema de conversión de texto en habla lleva a cabo la transformación de un texto escrito en su equivalente oral. En este capítulo se presenta, en primer lugar, la estructura general de un conversor de texto en habla (apartado 2), para exponer, a continuación, los diversos módulos que lo configuran (apartados 3 al 8). Para cada uno de ellos se intenta poner de relieve la información lingüística que requiere su desarrollo, teniendo en cuenta la bibliografía publicada y la experiencia de los autores (Llisterri, 2002). Finalmente, se ofrece una breve reflexión sobre el papel del lingüista en el ámbito de las tecnologías del habla.

Resumen:

El articulo presenta el proceso de conversión de texto en habla, explican los diferentes avances que se han realizado, el artículo no es precisamente un a investigación enfocado es un recopilado de diferentes técnicas, en las cuales se destacan las que utilizan sistemas expertos y redes neuronales como las que mas rendimiento tienen.

El artículo es de relevante al tema debido a que las técnicas que se utilizan son también utilizadas para realizar el proceso contrario.

Inproceedings
J. Orozco García, C.A.R.G.
Clasificación de Llanto del Bebé Utilizando una Red Neuronal de Gradiente Conjugado Escalado
1999

Abstract: El llanto es el único medio que un bebé tiene para comunicarse con el exterior. De acuerdo a los especialistas, en el llanto se refleja también el estado físico, patológico y/o anímico del bebé. El desarrollo de modelos dirigidos al estudio automático del llanto, permitirá proporcionar mejor atención al bebé para su cuidado. Este trabajo presenta el desarrollo de un sistema de reconocimiento automático del llanto del bebé, con ese objetivo en mente. Las características acústicas utilizadas se obtienen por medio de la técnica de predicción lineal y redes neuronales SCGB para clasificar entre los tipos de llantos de hambre, de dolor, de alguna patología y de placer. Se presentan los resultados iniciales obtenidos, con un número limitado de muestras, los cuales son muy promisorios.

Resumen:

El articulo resume proceso del Reconocimiento Automático del Llanto un Bebe, que es básicamente un problema de procesamiento de patrones, similar al Reconocimiento
Automático del Habla (RAH). Consisten tomar la onda de llanto del bebé como el patrón de entrada, y al final obtener el tipo de llanto o patología detectada en el bebé. El proceso de Reconocimiento Automático del Habla se hace en dos pasos. El primer paso es conocido como procesamiento de la señal, o extracción de características, mientras que el segundo se conoce como procesamiento o clasificación de patrones. En el análisis acústico, la señal de llanto del bebé es analizada para extraer las características más importantes en función del tiempo. Se efectúa el depuramiento de la señal tratando de eliminar la información irrelevante e indeseable
Como el ruido de fondo, distorsión del canal, y características particulares de la señal. Aunque los datos son reducidos al remover componentes repetitivos, la información relevante para la clasificación de patrones es conservada de una manera óptima. Algunas de las técnicas simples más usuales para el procesamiento de las señales son: coeficientes de predicción lineal, coeficientes, timbre, intensidad, análisis espectral, y bancos holmes entre otros. El conjunto de características obtenidas puede ser representado como un vector, y cada vector puede ser tomado como un patrón. El vector de características es comparado con el conocimiento que tiene la computadora. Por el lado de los métodos de reconocimiento de patrones, se han utilizado tradicionalmente cuatro enfoques principales: comparación de patrones, modelos estadísticos, sistemas basados en conocimientos y modelos conexionistas.

Inproceedings
S.Bhattacharya
Recognition of Voice signals for Oriya Language using wavelet Neural Network
2001

Abstract: Speech recognition is both speech oriented and speaker oriented. Both have fuzzy effect on them which adds to the hardness. During speech recognition, separation of the words and again redundancy of the voice responsible for the creation of words due to the vowels makes it difficult for analysis. We are trying with the wavelet neural network model to make an effective analysis for the recognition of speech signals. Here the main characteristics of speech/voice like frequency, intensity, accent and quality are analyzed for better output limiting the associated noise. Here we have used the concept of "wavelon" and "scalon". Using the control parameter of the network we have designed the wavelet network for two characters at the beginning and also we are extending for the rest of the character.

Resumen:

El artículo explica el procedimiento que se utiliza para el reconocimiento de palabras en un discurso continuo. Se implementa la técnica clásica de descomposición de la señal en este caso la voz en vectores, los cuales son insertados en la red la cual se encarga de la clasificación.

El trabajo presenta análisis de resultados importante ya que por medio de las pruebas realizadas concluyen que al segmentar el conjunto de palabras el margen de error de reconocimiento disminuye considerablemente, además este sistema se esta comparando con otros en la actualidad.

Inproceedings
John-Paul Hosom, R.A.C.
A DIPHONE-BASED DIGIT RECOGNITION SYSTEM USING NEURAL NETWORKS
1999

Abstract: In exploring new ways of looking at speech data, we have developed an alternative method of seg- mentation for training a neural-network-based digit-recognition system. Whereas previous methods segment the data into monophones, biphones, or triphones and train on each sub-phone unit in several broad-category contexts, our new method uses modi ed diphones to train on the regions of greatest spectral change as well as the regions of greatest stability. Although we account for regions of spectral stability, we do not require their presence in our word models. Empirical evidence for the advantage of this new method is seen by the 13% reduction in word-level error that was achieved on a test set of the OGI Numbers corpus. Comparison was

made to a baseline system that used conntext-independent monophones and context-dependent biphones and triphones.

Resumen:

El artículo presenta la investigación del proceso de reconocimiento de sílabas, en un ambiente dependiente (telefónico).

Se explica el proceso de descomposición de las señales en vectores, esto se realiza utilizando la teoría de tratamiento de señales, una vez que se ha descompuesto la señal esta se clasifica en vectores para facilitar el proceso de reconocimiento.

Mastersthesis
Fernandez, L.D.
Aportaciones a la Mejora de los Sistemas de Reconocimiento
Universida de de Vigo, ***2001***

Abstract: El Reconocimiento Automático de Voz (ASR Automatic Speech Recognition) es un campo de investigación de creciente relevancia que gana más adeptos. El desarrollo de mejores algoritmos y de modelados mías precisos, junto con la aparición de sistemas informáticos mías potentes y asequibles, posibilita la integración de los sistemas de dialogo hombre-maquina a través de la voz en numerosos ámbitos de la sociedad actual. Estos sistemas de dialogo permiten el acceso a una gran cantidad de información a través de una forma de comunicación tan natural como es el habla, facilitando un elevado numero de servicios interactivos utilizando el teléfono, la televisión o el ordenador como elementos de acceso. Los sistemas ASR se encuentran con una serie de dificultades cuando el canal de comunicación con el que van a trabajar no es predecible. Este problema es crucial en el desarrollo de aplicaciones factibles en dominios prometedores como son la telefonía y los coches. Los principales problemas encontrados se deben al locutor y a la tarea, al uso de micrófonos con diferentes características, a la calidad variable de los canales de transmisión, a la reverberación y ecos, a la distancia y dirección variable al micrófono introducida por el reconocimiento con manos-libres, y al ruido ambiente que distorsiona las señales de voz de entrada. El Reconocimiento Robusto de Voz trata con los desajustes entre entrenamiento y operación. Las técnicas mías recientes para reconocimiento robusto de voz se han enfocado principalmente en: 1) técnicas de PRE-procesado robusto de la señal de voz; y 2) compensación de características y modelos. Entre los métodos desarrollados para tratar con los desajustes entre datos de entrenamiento y de operación, las técnicas de adaptación/compensación están teniendo mucho interés debido a su capacidad para tratar con un amplio rango de variaciones de canal y ruido, junto con diferencias en locutores y estilos de habla. Sin embargo, mientras que un ser humano es capaz de adaptarse a una nueva voz con una cantidad mínima de datos de entrenamiento, la adaptación instantánea y no supervisada es todavía un gran reto para las máquinas. El interés surgido en la adaptación como procedimiento para resolver los

problemas de robustez proviene del buen nivel de prestaciones alcanzadas por los sistemas ASR y el esfuerzo dirigido hacia la inserción del ASR en aplicaciones del mundo real.

Phdthesis
TOLEDANO, D.T.
SEGMENTACIÓN Y ETIQUETADO FONÉTICOS AUTOMÁTICOS: Un Enfoque Basado en Modelos Ocultos de Markov y Refinamiento Posterior de las Fronteras Fonéticas
Señales, Sistemas y Radiocomunicaciones, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid., ***2000***

Abstract: El problema que tratamos de resolver en esta Tesis Doctoral es el problema de la segmentación y el etiquetado fonéticos automáticos. Este problema se puede definir de forma muy resumida como el problema de obtener, a partir de la señal de voz y de la trascripción ortográfica de la misma, la secuencia de fonemas pronunciada, así como las posiciones de las fronteras entre los mismos. Nótese que consideramos que la trascripción ortográfica (secuencia de palabras) es conocida a priori. Ésta es una suposición de orden práctico, ya que, por un lado, realizar una segmentación y un etiquetado fonético a partir de la voz exclusivamente es un problema todavía demasiado complejo como para ser realizado de forma automática con una precisión aceptable, y por otro, es muy frecuente disponer de la trascripción ortográfica de las pronunciaciones (y caso de no ser así, es relativamente poco costoso generarlas).

Mastersthesis
Ovalle, S.R.
Clasificación de Frases del Lenguaje Natural usando Redes Neuronales Recurrentes
Universidad Nacional de Colombia, ***2002***

Abstract: En este informe se considera la tarea de la clasificación de frases del lenguaje natural, utilizando redes neuronales recurrentes. Se obtiene como resultado la clasificación de las frases por su estado gramatical (gramaticalmente correctas o incorrectas). En este trabajo se logra en un porcentaje aceptable esta clasificación, utilizando como ejemplos para entrenar a la red neuronal recurrente, frases codificadas del lenguaje natural, basándose esta codificación en la teoría lingüística de Ligamiento. A partir del comportamiento de la red neuronal recurrente como sistema dinámico se hace una extracción de un autómata nito que corresponda a la gramática del lenguaje en cuestión. Para llevar a cabo estas tareas se consideró el desarrollo de una aplicación del sistema clasificador usando el algoritmo de Retropropagación a Través del Tiempo para el entrenamiento de la red neuronal y se utilizó el algoritmo de agrupamiento Gas Neuronal Creciente para extraer el autómata. XII

Phdthesis
GUARASA, J.M.
Arquitecturas y métodos en sistemas de reconocimiento automático de habla de gran vocabulario
Universidad politécnica de Madrid escuela técnica superior
De ingenieros de telecomunicación, ***2001***

Abstract: La tesis que se presenta en este documento, se enmarca en el área del Reconocimiento Automático de Habla y específicamente en el diseño de sistemas de reconocimiento de gran vocabulario. En todos los casos, la tecnología de base en lo que se refiere al modelado, la aportan los modelos ocultos de Markov que, hoy por hoy, representan el paradigma de modelado dominante. En concreto, se utilizarán técnicas de modelado discreto y semicontinuo, dependiente e independiente del contexto. En primer lugar, y a partir de una clasificación de alternativas arquitecturales en el diseño de sistemas de reconocimiento se hace un estudio teórico de la formulación del comportamiento de arquitecturas multi-módulo, tanto en coste computacional como en tasa de reconocimiento, definiendo una metodología de diseño para determinar la adecuación de módulos particulares de cara a su uso conjunto, que es validada con la experimentación correspondiente. Igualmente, se hace énfasis en el estudio y evaluación de algunas de las alternativas de compresión del espacio de búsqueda, estableciendo relaciones de compromiso entre coste y tasa, que es el binomio decisivo a la hora de abordar el diseño de sistemas en tiempo real. Se presentan estudios sobre distintas estrategias de organización del espacio de búsqueda orientadas a exploración y búsqueda con algoritmos de programación dinámica: árboles y grafos, deterministas y no deterministas, proponiendo soluciones prometedoras para incrementar la tasa de inclusión obtenible sobre estructuras de grafo (en las que la compresión del espacio de búsqueda produce peores resultados que con la búsqueda lineal o en árbol). Especialmente importante es el trabajo sobre estimación de listas variables de preselección, analizando métodos paramétricos y no paramétricos, centrándonos en el uso de redes neuronales como mecanismo estimador. Se ha propuesto una metodología de selección de parámetros de entrada, topologías y métodos de codificación, en base a su potencia discriminativa en una tarea simplificada. Dicha propuesta que ha sido ampliamente evaluada y comparada con el enfoque tradicional de uso de listas fijas, mostrando la consistente mejora tanto en tasa como en coste computacional conseguible con el uso de redes neuronales. Dicho estudio sobre listas variables ha sido extendido de forma natural al problema de estimación de fiabilidad de hipótesis, habiéndose aprovechando estos resultados, de nuevo, para la estimación de longitudes de listas, obteniendo también buenos resultados. En lo que respecta al repertorio de unidades de reconocimiento y a la composición de los diccionarios usados (en cuanto al uso de múltiples pronunciaciones), se aplican, evalúan y comparan métodos dirigidos por datos y basados en conocimiento. En el apartado de introducción de variantes de pronunciación se ha discutido ampliamente la problemática de contar con bases de datos representativas y haciendo énfasis en la importancia de atender y evaluar las mejoras marginales obtenidas con algunos de estos métodos. La evaluación de los resultados es

planteada cuidadosamente, sobre dos tareas radicalmente distintas: habla telefónica independiente del locutor y habla aislada dependiente, ambas usando gran vocabulario (hasta 10000 palabras), lo que permite obtener conclusiones y claves de diseño para cada una de ellas, con lo que se consigue una generalización más fundamentada de sus bondades o perjuicios. En este sentido se aplican análisis de validez y relevancia estadística que pongan en su justo sitio las mejoras o degradaciones observadas. En los procesos de evaluación se han propuesto nuevas métricas y mecanismos originales de comparación.