Diss. ETH Nr. 8742

Beiträge zum Problem der textabhängigen Sprecherverifikation

ABHANDLUNG

zur Erlangung des Titels eines

DOKTORS DER TECHNISCHEN WISSENSCHAFTEN

der

EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE

ZÜRICH

vorgelegt von

Carlo Bernasconi

dipl. El.-Ing. ETH

geboren am 10. Mai 1955

von Castel San Pietro TI

angenommen auf Antrag von

Prof. Dr. W. Guggenbühl, Referent

Prof. Dr. G. Moschytz, Korreferent

Zürich, den 15. Dezember 1988

i. V. Guggenbühl

Zusammenfassung

Es wird ein leistungsfähiges textabhängiges Sprecherverifikationsverfahren beschrieben, welches eine sehr hohe Entscheidungssicherheit zu erreichen erlaubt. Seine Hauptkomponenten sind die Vorverarbeitung, welche Digitalisierung, Anfangs- und Endpunktdetektion und Analyse umfasst, die Zeitnormalisation und die Entscheidungsstufe. Nach jedem erfolgreichen Verifikationsversuch findet eine Nachführung der Referenzkontur statt.

Zur Analyse wird die LPC(linear prediction coding)-Methode verwendet. Als Merkmale des Sprachsignals werden einzig die AR(autoregressive)-Parameter eingesetzt (keine Information über Stimmbändergrundfrequenz oder Signalleistung). Die Zeitnormalisation beruht auf einem speziellen DTW(dynamic time warping)-Algorithmus, welcher UELM(unconstrained endpoints, local minimum)-Verfahren genannt wird. Die Bildung der Referenzkontur eines gegebenen Sprechers erfolgt ausgehend von mehreren Trainingssätzen aus einer einzigen Aufnahmesession.

Jede Teiloperation der Verifikationsprozedur wurde nach ausführlichen Untersuchungen optimal ausgelegt. Ein automatisches Anfangs- und Endpunktdetektionsverfahren, welches der Intensität des Hintergrundlärms Rechnung trägt, wurde zuerst entwickelt und erfolgreich ausgetestet.

Anschliessend wurde die für die Entscheidungssicherheit wichtige Zeitnormalisation eingehend analysiert, um die Ursache fataler Normalisationsfehler unter Hypothese H_0 (der Testsatz gehört zum fraglichen Sprecher) zu eliminieren. Dabei stellte sich heraus, dass lange Sprechpausen für derartige Fehler verantwortlich sind. Das entwickelte

automatische Pausenextraktionsverfahren liefert hier eine befriedigende Lösung.

Ferner wurde die Diskriminierungsfähigkeit der cepstralen Koeffizienten und anderer daraus hervorgehender Parametersätze wie normalisierte cepstrale Koeffizienten oder orthogonale Koeffizienten erster Ordnung, welche aus der Entwicklung der cepstralen Konturen in orthogonalen Polynomen resultieren, untersucht. Die signifikantesten Ergebnisse sind hier einerseits die erhöhte Diskriminierung mittels Elimination der zeitinvarianten spektralen Komponente der Sprachprobe, anderseits die ausgezeichneten diskriminierenden Eigenschaften der orthogonalen Koeffizienten erster Ordnung.

Zwei Methoden zur Referenzbildung, welche sich etwa gleichwertig zeigten, wurden ausgelegt. Wichtig in diesem Zusammenhang ist, dass für eine repräsentative Referenzkontur mindestens vier Trainingskonturen aus derselben Aufnahmesession eingesetzt werden müssen.

Die verwendete Referenznachführungsmethode bewirkt kurzfristig eine deutliche Erhöhung der Diskriminierung. Mit dem vorhandenen Sprachmaterial konnte aber nur eine beschränkte Untersuchung über das langfristige Verhalten durchgeführt werden. Umfangreiche Versuche zeigten schliesslich, dass eine Nachführung der zeitlichen Struktur (Dauer der phonetischen Ereignisse) im Referenzsatz keine Vorteile bringt, eine Anpassung lediglich der spektralen Eigenschaften ist folglich ausreichend.

Das entwickelte Verifikationsverfahren wurde anhand einer Testpopulation von 26 männlichen Sprechern mit insgesamt über 1300 Sprachproben getestet. Diese wurden über eine Zeitspanne von mehr als drei Jahren, stets unter denselben guten akustischen Bedingungen (keine telephonische Übermittlung der Sprachproben) aufgenommen.

Umfangreiche Verifikationsexperimenten (jedes davon mit insgesamt ca. 3000 Verifikationsversuchen) ergaben sehr niedrige mittlere Fehlerraten (um 0.1 %). Sie zeigten ferner, dass es nicht nötig ist, die Entscheidungsschwelle individuell zu optimieren. Aus den experimentellen Ergebnissen kann man schliessen, dass mit dem entwickelten Sprecherverifikationsverfahren eine hohe Entscheidungssicherheit erreicht werden kann.

Abstract

In the present work a text dependent speaker verification procedure based on classical pattern recognition methods is developed. Its main components are preprocessing (which includes A/D-conversion, end-point detection, and signal analysis), time alignment, and decision step. Reference contours are updated after every successful verification.

LPC(linear prediction coding)-analysis of the speech wave is performed but no information about pitch and gain is taken into account. A special DTW(dynamic time warping)-algorithm, the UELM(unconstrained endpoints, local minimum)-method is employed for time alignment. Reference contours for a given speaker are formed using training sentences from a single recording session.

An efficient algorithm for endpoints detection has been first developed and successfully tested. This algorithm estimates the background noise power intensity to get a reliable voice/silence decision.

Failures of the time alignment procedure, especially for the critical cases where reference contour and test utterance belong to the same speaker (hypothesis H_0), have been investigated. The main result is, that fatal alignment errors can be attributed to long silence intervals in the speech signal. The detection and elimination of such intervals is a very powerful solution of this problem.

Some LPC-cepstrum based features of the speech wave, like the normalized cepstral coefficients and the first order orthogonal coefficients, resulting from the orthogonal polynomial expansion of the cepstral time functions, have been investigated in order to find those with the better properties for speaker verification. The most significant results are here that a spectral normalization produces a consistent inprovement in the

decision reliability and that first order orthogonal coefficients perform as well as normalized cepstral coefficients.

Both reference building methods investigated perform almost equally well. The most important result in this context is that at least four training utterances from the same recording session must be employed in order to build a representative reference.

Reference updating produces a clear improvement of the verification reliability after the first few successful verification attempts. This is due to the fact that the updating method takes new speaker related information into the references. Moreover it could be shown that updating the dynamic (duration) of the phonetical events in the reference contour is not required, only updating the spectral features is needed.

The developed verification algorithm has been extensively tested on a population of 26 male speakers with a total of about 1300 sentence long utterances, recorded over a period of over three years. All recordings took place always under the same good acoustic conditions (clear speech, no telephone speech has been included).

The average error rates obtained over about 3000 verification attempts were very low (around 0.1 %), even using a common decision threshold for all speakers. A separate optimal choice of the threshold for every speaker is therefore not required. Considering these results, a very high reliability of the developed verification procedure using clear speech can be expected.

Riassunto

Nel presente lavoro viene trattato un metodo di verifica dell'identità del locutore che si basa su frasi di prova a testo fisso. Le sue componenti principali sono: la fase preparatoria (digitalizzazione, soppressione delle pause prima e dopo la frase di prova e analisi), l'allineamento nel tempo e la fase di decisione. I contorni di referenza dei locutori vengono aggiornati dopo ogni verifica.

L'analisi viene eseguita secondo il metodo LPC (linear prediction coding). Per caratterizzare la voce ci si serve solo dei vettori dei coefficienti autoregressivi (nessuna informazione sul volume e sulla frequenza delle corde vocali viene utilizzata). L'allineamento nel tempo avviene tramite un algoritmo DTW (dynamic time warping) chiamato metodo UELM (unconstrained endpoints, local minimum). Per la formazione dei contorni di referenza di un dato locutore ci si serve di frasi provenienti da un'unica seduta di registrazione.

Un sistema automatico per individuare l'inizio e la fine della frase di prove nel segnale acustico è stato dapprima sviluppato e provato con successo. Esso si adatta all'intensità dei rumori di fondo.

L'algoritmo che esegue l'allineamento temporale è stato analizzato attentamente con lo scopo di individuare le cause di errori gravi, i quali, sotto l'ipotesi H_0 (la voce appartiene veramente alla persona in causa), possono precludere una corretta verifica. Dalle analisi effettuate risulta che pause di dicitura troppo lunghe sono quasi sempre la causa di suddetti errori. Il procedimento automatico di eliminazione delle pause sviluppato in proposito permette di ottenere una soluzione soddisfacente del problema.

Le capacità discriminatorie dei coefficienti del cepstro e di altri vet-

tori da essi derivanti come i coefficienti del cepstro normalizzati oppure i coefficienti ortogonali di primo ordine che si ottengono dallo sviluppo delle successioni dei coefficienti del cepstro in polinomi ortogonali, sono state studiate a fondo. I risultati più importanti sono da una parte il miglioramento dell'affidabilità ottenuto tramite l'eliminazione dai segnali acustici della loro componente spettrale costante nel tempo, d'altra parte le eccellenti qualità discriminatorie dei coefficienti ortogonali di primo ordine.

I due procedimenti per la formazione dei contorni di referenza studiati hanno dato praticamente lo stesso risultato. In questo contesto è importante il fatto che per la formazione di una referenza sufficientemente rappresentativa sono necessarie almeno quattro frasi di allenamento.

Il metodo impiegato per l'aggiornamento della referenza produce, a corto termine, un notevole aumento dell'affidabilità della verifica dovuto al coinvolgimento di nuove frasi nella referenza stessa. Simulazioni esaurienti hanno inoltre provato che non è necessario aggiornare anche la struttura temporale (durata degli eventi fonetici) nella frase di referenza ma che un aggiornamento delle caratteristiche spettrali è sufficiente.

Il metodo di verifica sviluppato è stato provato su una popolazione di 26 persone, tutte di sesso maschile, con l'impiego di più di 1300 frasi. Le registrazioni, distribuite su oltre tre anni, sono state eseguite sempre sotto le stesse condizioni acustiche favorevoli (poco rumore di fondo, nessuna trasmissione della voce per telefono).

Gli esperimenti effettuati, comprendenti ognuno circa 3000 tentativi di verifica, hanno dato dei tassi d'errore molto bassi (attorno al 0.1 %). Inoltre hanno mostrato che non è necessario fissare soglie di decisione individuali, scelte in modo ottimale per ogni persona. I risultati ottenuti permettono senz'altro di concludere che, con il metodo sviluppato, è possibile ottenere un'alta affidabilità nella verifica del locutore.

Résumé

Le présent travail traite une méthode dépendente du texte pour la vérification automatique de l'identité du locuteur. Cette méthode consiste du prétraitement (qui comprend la digitalisation, la détection des éxtremités de la phrase d'essai dans le signal de la parole et son analyse), de l'alignement temporel et de la phase de décision. La phrase de référence est mise à jour après chaque vérification réussite.

L'analyse du signal se fait par la méthode LPC (linear prediction coding), aucune information au sujet du pitch ou du gain n'est cependant prise en considération. Un algorithme DTW (dynamic time warping) spécial, appelé UELM (unconstrained endpoints local minimum), est utilisé pour l'alignement temporel. La phrase de référence pour un locuteur donné est formée en partant de phrases d'entraînement enregistrées pendant une seule séance de registration.

D'abord un algorithme robuste et éfficace pour la détection des éxtremités de la phrase d'essai a été développé et employé avec succes. Cet algorithme utilise une estimation de la puissance du bruit de fond pour distinguer entre parole et silence.

Les défaillances de la procédure d'alignement temporel ont été étudiées attentivement, surtout lorsque référence et phrase d'essai appartiennent au même locuteur (hypothèse H_0). Il s'avérat que la majorité de ces fautes sont provoquées par des pauses de diction de longueur excessive. L'élimination automatique de ces pauses représente une solution efficace du problème.

Les proprietés discriminatoires de coefficients basées sur le cepstrum LPC, comme les coefficients cepstraux normalisés ou les coefficients orthogonaux de premier ordre, que l'on obtient en développant les sui-

tes temporelles des coefficients cepstraux en polynomes orthogonaux, ont été étudiées. Ces études ont donné deux résultats importants: premier, que la fiabilité de la verification augmente en éliminant de la phrase d'essai la partie du spectrum qui reste constante dans le temps, deuxième, que des coefficients orthogonaux de premier ordre présentent des excellents proprietés discriminatoires, comparable avec ceux des coefficients cepstraux normalisés.

Les deux méthodes de formation des phrases de réference qu'on a mis à point, donnent à peu près les mêmes résultats. Au moins quatre phrases d'entraînement sont nécessaire à la formation d'une référence assez représentative.

La mise à jour de la référence provoque, à court terme, une amélioration de celle ci qui comporte aussi une augmentation de la fiabilité de la verification. Nous montrons aussi que la mise à jour des caractéristiques spectrales seules est suffisante, la dynamique (durée) des évents phonétiques n'ayant pas besoin d'être adaptée.

La performance du système développé a été mesurée sur une population de 26 personnes de sexe masculin avec un total d'environ 1300 phrases d'essai. Toutes ces phrases ont été enregistrées pendant plus de trois ans, toujours sous les mêmes conditions acoustiques assez favorables (pas de bruit de fond notable et pas de transmission de la voix par téléphone).

Expériments, chaqun avec environ 3000 essais, ont donné des taux d'erreur très bas (près de 0.1 %), même en utilisant un seul seuil de décision pour tous les locuteurs. Le choix d'un seuil individuel n'est ainsi pas nécessaire. D'après ces résultats on peut affirmer qu'avec la méthode développée dans le cadre de ce projet il est possible d'obtenir une très haute fiabilité dans la vérification de l'identité du locuteur.