(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2002 (27.12.2002)

PCT

(10) International Publication Number
WO 02/103680 A2

(51) International Patent Classification[7]: G10L 17/00

(21) International Application Number: PCT/GB02/02726

(22) International Filing Date: 13 June 2002 (13.06.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0114866.7 19 June 2001 (19.06.2001) GB
60/302,501 2 July 2001 (02.07.2001) US

(71) Applicant *(for all designated States except US)*: **SECURIVOX LTD** [GB/GB]; 336 Perth Road, Dundee DD2 1EQ (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **SAPELUK, Andrew, Thomas** [GB/GB]; 4 Burn Street, Dundee DD3 0LA (GB).

(74) Agent: **MURGITROYD & COMPANY**; Scotland House, 165-169 Scotland Street, Glasgow G5 8PL (GB).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/103680 A2

(54) Title: "SPEAKER RECOGNITION SYSTEMS"

(57) Abstract: Speaker recognition (identification and/or verification) methods and systems, in which speech models for enrolled speakers consist of sets of feature vectors representing the smoothed frequency spectrum of each of a plurality of frames and a clustering algorithm is applied to the feature vectors of the frames to obtain a reduced data set representing the original speech sample, and wherein the adjacent frames are overlapped by at least 80 %. Speech models of this type model the static components of the speech sample and exhibit temporal independence. An identifier strategy is employed in which modelling and classification processes are selected to give a false rejection rate substantially equal to zero. Each enrolled speaker is associated with a cohort of a predetermined number of other enrolled speakers and a test sample is always matched with either the claimed identity or one of its associated cohort. This makes the overall error rate of the system dependent only on the false acceptance rate, which is determined by the cohort size. The false error rate is further reduced by use of multiple parallel modelling and/or classification processes. Speech models are normalised prior to classification using a normalisation model derived from either the test speech sample or one of the enrolled speaker samples (most preferably from the claimed identity enrolment sample).

"Speaker Recognition Systems"

The present invention relates to systems, methods and apparatus for performing speaker recognition.

Speaker recognition encompasses the related fields of speaker verification and speaker identification. The main objective is to confirm the claimed identity of a speaker from his/her utterances, known as verification, or to recognise the speaker from his/her utterances, known as identification. Both use a person's voice as a biometric measure and assume a unique relationship between the utterance and the person producing the utterance. This unique relationship makes both verification and identification possible. Speaker recognition technology analyses a test utterance and compares it to a known template or model for the person being recognised or verified. The effectiveness of the system is dependent on the quality of the algorithms used in the process.

Speaker recognition systems have many possible applications. In accordance with a further aspect of the present invention, speaker recognition technology may be used to permanently mark an electronic document with a biometric print for every person who views or edits the content. This produces an audit trail identifying all of the users and the times of access and modification. As the user mark is biometric it is very difficult for the user to dispute the authenticity of the mark.

Other biometric measures may provide the basis for possible recognition systems, such as iris scanning, finger printing and facial features. These measures all require additional hardware for recording whereas speaker recognition can be used with any voice input such as over a telephone line or using a standard multi-media personal computer with no modification. The techniques can be used in conjunction with other security measures and other biometrics for increased security. From the point of view of a user the operation of the system is very simple.

For example, when an on-line document is requested the person requiring access will be asked to give a sample of their speech. This will be a simple prompt from the client software *'please say this phrase..."* or something similar. The phrase uttered will then be sent to a database server or to a speech recognition server, via any data network such as an intranet, to be associated with the document and

stored as the key used to activate the document at that particular time. A permanent record for a document can therefore be produced, over time, providing an audit trail for the document. The speaker authentication server may maintain a set of templates (models) for all currently enrolled persons and a historical record of previously enrolled persons.

Speaker recognition systems rely on extracting some unique features from a person's speech. This in turn depends on the manner in which human speech is produced using the vocal tract and the nasal tract. For practical purposes, the vocal tract and nasal tract can be regarded as two connected pipes, which can resonate in a manner similar to a musical instrument. The resonances produced depend on the diameter and length of the pipes. In the human speech production mechanism, these diameters and to some extent the length of the pipe sections can be modified by the articulators, typically the positions of the tongue, the jaw, the lips and the soft palate (velum). These resonances in the spectrum are called the formant frequencies. There are normally around four formant frequencies in a typical voice spectrum.

As with musical instruments, sound will only be produced when a constriction of the airflow occurs causing either vibration or turbulence. In human speech, the major vibrations occur when the constriction occurs at the glottis (vocal cords).

When this happens, voiced speech is produced, typically vowel-like sounds. When the constriction is in the mouth, caused by the tongue or teeth, a turbulence is produced, (a hissing type of sound) and the speech produced is called a fricative, typified by "s", "sh", "th" etc. From an engineering point of view, this is similar to a source signal (the result of the constriction) being applied to a filter which has the general characteristics (i.e. the same resonances) of the vocal tract and the resulting output signal is the speech sound. True speech is produced by dynamically varying the positions of the articulators.

All existing speaker recognition systems perform similar computational steps. They operate by creating a template or model for an enrolled speaker. The model is created by two main steps applied to a speech sample, namely spectral analysis and statistical analysis. Subsequent recognition of an input speech sample is performed by modelling the input sample (test utterance) in the same way as during speaker enrolment, and pattern/classification matching of the input model against a database of enrolled speakers. Existing systems vary in the approach taken when performing some or all of these steps. In conventional (industry standard) systems, the spectral analysis is either Linear Predictive Coding (LPC)/Cepstral analysis ("LPCC") or FFT/sub-banding. This is followed by a statistical analysis technique, usually a technique called Hidden Markov

Modelling (HMM), and the classification step is a combination of a match against the claimed speaker model and against an "impostor cohort" or "world model" (i.e. a set of other speaker models).

To allow efficient processing of speech samples, all speaker recognition systems use time slices called frames, where the utterance is split into frames and each frame is processed in turn. Frames may or may not be of equal size and may or may not overlap. An example of a typical time signal representation of a speech utterance divided into frames is illustrated in Fig. 1 of the accompanying drawings. A generic speaker recognition system is shown in block diagram form in Fig. 2, illustrating a test utterance being processed, through an input filter 10, a spectral analysis (LPCC) stage 12 and a statistical analysis (HMM) stage 14, followed by score normalisation and speaker classification 16, by thresholding, employing a database 18 of speaker models (enrolled speaker data-set), before generating a decision as to the identity of the speaker (identification) or the veracity of the speaker's claimed identity (verification).

Such systems have a number of disadvantages or limitations. Firstly, conventional spectral analysis techniques produce a limited and incomplete feature set and therefore poor modelling. Secondly, HMM techniques are "black-box" methods, which combine good performance with relative ease of use, but at the expense of transparency. The relative

importance of features extracted by the technique are not visible to the designer. Thirdly, the nature of the HMM models do not allow model-against-model comparisons to be made effectively. Accordingly, important structural detail contained within the enrolled speaker data-set cannot be analysed and used effectively to improve system performance. Fourthly, HMM technology uses temporal information to construct the model and is therefore vulnerable to mimics, who impersonate others' voices by temporal variations in pitch etc. Fifthly, the world model/impostor cohort employed by the system cannot easily be optimised for the purpose of testing an utterance by a claimed speaker.

The performance of a speaker recognition system relies on the fact that when a true speaker utterance is tested against a model for that speaker it will produce a score, which is lower than a score that is produced when an impostor utterance is tested against the same model. This allows an accept/reject threshold to be set. Consecutive tests by the true speaker will not produce identical scores. Rather, the scores will form a statistical distribution. However, the mean of the true speaker distribution will be considerably lower than the means of impostor distributions tested against the same model. This is illustrated in Fig. 3, where 25 scores are plotted for each of eight speakers, speaker 1 being the true speaker. It can be seen from Fig. 3 that the scores of some speakers are

closer to the true speaker than others and can be problematic.

The present invention relates to improved speaker recognition methods and systems which provide improved performance in comparison with conventional systems. In various aspects, the invention provides improvements including but not limited to: improved spectral analysis, transparency in its statistical analysis, improved modelling, models that can be compared allowing the data-set structure to be analysed and used to improve system performance, improved classification methods and the use of statistically independent/partially independent parallel processes to improve system performance.

The invention further embraces computer programs for implementing the methods and systems of the invention, data carriers and storage media encoded with such programs, data processing devices and systems adapted to implement the methods and systems, and data processing systems and devices incorporating the methods and systems.

The various aspects and preferred features of the invention are defined in the Claims appended hereto.

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Fig. 1 is a time signal representation of an example of a speech utterance divided into frames;

Fig. 2 is a block diagram of a generic, prior art speaker recognition system;

Fig. 3 is a plot of speaker recognition score distributions for a number of speakers tested against one of the speakers, obtained using a conventional speaker recognition system;

Fig. 4 is a block diagram illustrating a first embodiment of the present invention;

Fig. 5 is a block diagram illustrating a second embodiment of the present invention;

Fig. 6 is a block diagram illustrating a third embodiment of the present invention;

Fig. 7 is a block diagram illustrating a further embodiment of a speaker recognition system in accordance with the present invention;

Fig. 8(a) is a time signal representation of an example of a speech utterance divided into frames and Fig. 8(b) shows the corresponding frequency spectrum and smoothed frequency spectrum of one frame thereof;

Fig. 9 illustrates the differences between the frequency spectra of two mis-aligned frames;

Fig. 10 shows the distribution of accumulated frame scores plotted against their frequency of occurrence;

Fig. 11(a) shows the same accumulated score distributions as Fig. 3 for comparison with Fig. 11(b), which shows corresponding accumulated score distributions obtained using a speaker recognition system in accordance with the present invention;

Fig. 12 illustrates the results of model against model comparisons as compared with actual test scores, obtained using a system in accordance with the present invention;

Fig. 13 illustrates the distribution of speaker models used by a system in accordance with the present invention in a two-dimensional representation of a multi-dimensional dataspace;

Fig. 14 illustrates the use of an impostor cohort as used in a system in accordance with the present invention;

Fig. 15 is a block diagram illustrating a normalisation process in accordance with one aspect of the present invention;

Fig. 16 is a block diagram illustrating an example of wide area user authentication system in accordance with the present invention;

Fig. 17 is a block diagram illustrating the corruption of a speech signal by various noise sources and channel characteristics in the input channel of a speaker recognition system;

Figs. 18 and 19 illustrate the effects of noise and channel characteristics on test utterances and enrolment models in a speaker recognition system; and

Fig. 20 illustrates a channel normalisation method in accordance with one aspect of the present invention.

The present invention includes a number of aspects and features which may be combined in a variety of ways in order to provide improved speaker

recognition (verification and/or identification) systems. Certain aspects of the invention are concerned with the manner in which speech samples are modelled during speaker enrolment and during subsequent recognition of input speech samples. Other aspects are concerned with the manner in which input speech models are classified in order to reach a decision regarding the identity of the speaker. A further aspect is concerned with normalising speech signals input to speaker recognition systems (channel normalisation). Still further aspects concern applications of speaker recognition systems.

Referring now to the drawings, Figs. 4 to 6 illustrate the basic architectures used in systems embodying various aspects of the invention. It will be understood that the inputs to all of the embodiments of the invention described herein are digital signals comprising speech samples which have previously been digitised by any suitable means (not shown), and all of the filters and other modules referred to are digital.

In Fig. 4, a speech sample is input to the system via a channel normalisation module 200 and a filter 24. Instead of or in addition to this "front-end" normalisation, channel normalisation may be performed at a later stage of processing the speech sample, as shall be discussed further below. The sample would be divided into a series of frames prior to being input to the filter 24 or at some other point prior to feature extraction. In some

embodiments, as discussed further below, a noise signal 206 may be added to the filtered signal (or could be added prior to the filter 24). The sample data are input to a modelling (feature extraction) module 202, which includes a spectral analysis module 26 and (at least in the case of speech sample data being processed for enrolment purposes) a statistical analysis module 28. The model (feature set) output from the modelling module 202 comprises a set of coefficients representing the smoothed frequency spectrum of the input speech sample. During enrolment of a speaker, the model is added to a database of enrolled speakers (not shown). During recognition of an input speech sample, the model (feature set) is input to a classification module 110, which compares the model (feature set) with models selected from the database of enrolled speakers. On the basis of this comparison, a decision is reached at 204 so as to identify the speaker or to verify the claimed identity of the speaker. The channel normalisation of the input sample and the addition of the noise signal 206 comprise aspects of the invention, as shall be described in more detail below, and are preferred features of all implementations of the invention. In some embodiments, channel normalisation may be applied following spectral analysis 26 or during the classification process, rather than being applied to the input speech sample prior to processing as shown in Figs. 4 to 6. Novel aspects of the modelling and classification processes in accordance with other

aspects of the invention will also be described in more detail below.

Other aspects of the invention involve various types of parallelism in the processing of speech samples for enrolment and/or recognition.

In Fig. 5, the basic operation of the system is the same as in Fig. 4, except that the output from the modelling module 202 is input to multiple, parallel classification processes 110a, 110b ... 110n, and the outputs from the multiple classification processes are combined in order to reach a final decision, as shall be described in more detail below. In Fig. 6, the basic operation of the system is also the same as in Fig. 4, except that the input sample is processed by multiple, parallel modelling processes 202a, 202b ... 202n (typically providing slightly different feature extraction/modelling as described further below), possibly via multiple filters 24a, 24b ... 24n (in this case the noise signal 206 is shown being added to the input signal upstream of the filters 24a, 24b ... 24n), and the outputs from the multiple modelling processes are input to the classification module 110, as shall also be described in more detail below. These types of multiple parallel modelling processes are preferably applied to both enrolment sample data and test sample data.

Multiple parallel modelling processes may also be combined with multiple parallel classification

processes; e.g. the input to each of the parallel classification processes 110a-n in Fig. 5 could be the output from multiple parallel modelling processes as shown in Fig. 6.

Various aspects of the invention will now be described in more detail by reference to the modelling, classification and normalisation processes indicated in Figs. 4 to 6.

MODELLING

The spectral analysis modules 26, 26a-n may apply similar spectral analysis methods to those used in conventional speaker recognition systems. Preferably, the spectral analysis applied by the modules 26a-n is of a type that, for each frame of the sample data, extracts a set of feature vectors (coefficients) representing the smoothed frequency spectrum of the frame. This preferably comprises LPC/Cepstral (LPCC) modelling, producing an increased feature set which models the finer detail of the spectra, but may include variants such as delta cepstral or emphasis/de-emphasis of selected coefficients based on a weighting scheme. Similar coefficients may alternatively be obtained by other means such as or Fast Fourier Transform [FFT] or by use of a filter bank.

The complete sample is represented by a matrix consisting of one row of coefficients for each frame of the sample. For the purposes of the preferred

embodiments of the present invention, these matrices will each have a size of the order of 1000 (frames) x 24 (coefficients). In conventional systems, a single first matrix of this type, representing the complete original signal, would be subject to statistical analysis such as HMM.

As will be understood by those skilled in the art, the LP transform effectively produces a set of filter coefficients representing the smoothed frequency spectrum for each frame of the test utterance. The LP filter coefficients are related to Z-plane poles. The Cepstral transform has the effect of compressing the dynamic range of the smoothed spectrum, de-emphasising the LP poles by moving them closer to the Z-plane origin (away from the real frequency axis at $z=e^{jw}$). The Cepstral transform uses a log function for this purpose. It will be understood that other similar or equivalent techniques could be used in the spectral analysis of the speech sample in order to obtain a smoothed frequency spectrum and to de-emphasise the poles thereof. This de-emphasis produces a set of coefficients which when transformed back into the time domain are less dynamic and more well balanced (the cepstral coefficients are akin to a time signal or impulse response of the LP filter with de-emphasised poles). The log function also transforms multiplicative processes into additive processes.

The model derived from the speech sample may be regarded as a set of feature vectors based on the

frequency content of the sample signal. When a feature vector based on frequency content is extracted from a signal, the order of the vector is important. If the order is too low then some important information may not be modelled. To avoid this, the order of the feature extractor (e.g. the number of poles of an LP filter) may be selected to be greater than the expected order. However, this in itself causes problems. Poles which match resonances in the signal give good results, whilst the other resulting coefficients of the feature vector will model spurious aspects of the signal. Accordingly, when this vector is compared with another model or reference, the distance measure computed may be unduly influenced by the values of those coefficients which are modelling spurious aspects of the signal. The distance measure (score) which is returned will thus be inaccurate, possibly giving a poor score for a frame which in reality is a good match.

In accordance with one aspect of the invention, this problem can be obviated or mitigated by adding a noise signal n(t) (206 in Figs. 4 - 6) having known characteristics to the speech signal s(t) before the signal is input to the modelling process (i.e. the input signal = s(t)+n(t)). The same noise signal would be used during enrolment of speakers and in subsequent use of the system. The addition of the known noise signal has the effect of forcing the "extra" coefficients (above the number actually required) to model a known function and hence to

give consistent results which are less problematic during model/test vector comparison. This is particularly relevant for suppressing the effect of noise (channel noise and other noise) during "silences" in the speech sample data. This problem may also be addressed as a consequence of the use of massively overlapping sample frames discussed below.

As previously mentioned, in order to allow efficient processing of speech samples all speaker recognition systems use time slices called frames, so that the utterance is split into a sequence of frames and each frame is processed in turn. The frames may or may not be of equal size and they may overlap. Models generated by speaker recognition systems thus comprise a plurality of feature sets (vectors corresponding to sets of coefficients) representing a plurality of frames. When models are compared in conventional speaker recognition systems it is necessary to align corresponding frames of the respective models. Different utterances of a given phrase will never be exactly the same length, even when spoken by the same person. Accordingly, a difficulty exists in correctly aligning frames for comparison.

Conventional systems convert the frames into a spectral or smoothed spectral equivalent as shown in Figs. 8(a) (showing a time signal divided into frames) and 8(b) (showing the corresponding frequency spectrum and smoothed frequency spectrum of one of the frames of Fig. 8(a)). The systems

then perform further transformations and analysis (such as Cepstral transformation, Vector Quantisation, Hidden Markov Modelling (HMM) and Dynamic Time Warping (DTW)) to obtain the desired result. Frame boundaries can be allocated in many ways, but are usually measured from an arbitrary starting point estimated to be the starting point of the useful speech signal. To compensate for this arbitrary starting point, and also to compensate for the natural variation in the length of similar sounds, techniques such as HMM and DTW are used when comparing two or more utterances such as when building models or when comparing models with test utterances. The HMM/DTW compensation is generally done at a point in the system following spectral analysis, using whatever coefficient set is used to represent the content of a frame, and does not refer to the original time signal. The alignment precision is thus limited to the size of a frame. In addition, these techniques assume that the alignment of a particular frame will be within a fixed region of an utterance which is within a few frames of where it is expected to lie. This introduces a temporal element to the system as the estimated alignment of the current frame depends on the alignment of previous frames, and the alignment of subsequent frames depends on the alignment of the present frame. In practice, this means that a particular frame, such as a frame which exists 200 ms into an utterance, will in general only be compared with other frames in the 200 ms region of the model or of other utterances being used to

construct a model. This approach derives from speech recognition methods (e.g. speech-to-text conversion), where it is used to estimate a phonetic sequence from a series of frames. The present applicants believe that this approach is inappropriate for speaker recognition, for the following reasons.

A. Most seriously, the conventional approach provides only crude alignment of frames. The arbitrary allocation of starting points means that it will generally not be possible to obtain accurate alignment of the starting points of two respective frames, so that even two frames giving a "best match" may have significantly different spectral characteristics, as illustrated in Fig. 9.

B. Secondly, the conventional approach relies on the temporal sequence of the frames and bases speaker verification on spectral characteristics derived from temporally adjacent frames.

In accordance with a further aspect of the invention, the present enrolment modelling process involves the use of very large frame overlaps, akin to convolution, to avoid problems arising from frame alignment between models (discussed at A. above) and to improve the quality of the model obtained. This technique is applied during speaker enrolment in order to obtain a model, preferably based on repeated utterances of the enrolment phrase. By massively overlapping the frames, the resulting

model effectively approaches a model of all possible alignments, with relatively small differences between adjacent frames, thereby providing good modelling of patterns. Preferably, the frame overlap is selected to be at least 80%, more preferably it is in the range 80% to 90%, and may be as high as 95%.

The frames are transformed into representative coefficients using the LPCC transformation as described above, so that each utterance employed in the reference model generated by the enrolment process is represented by a matrix (typically having a size of the order of 1000 frames by 24 coefficients as previously described). There might typically be ten such matrices representing ten utterances. A clustering or averaging technique such as Vector Quantisation (described further below) is then used to reduce the data to produce the reference model for the speaker. This model does not depend on the temporal order of the frames, addressing the problems described at B. above.

Preferred embodiments of the present invention combine the massive overlapping of frames described above with Vector Quantisation or the like as described below. This provides a mode of operation which is quite different from conventional HMM/DTW systems. In such conventional systems, all frames are considered equally valid and are used to derive a final "score" for thresholding into a yes/no decision, generally by accumulating scores derived

by comparing and matching individual frames. The validity of the scores obtained is limited by the accuracy of the frame alignments.

In accordance with this aspect of the present invention, the reference (enrolment) models represent a large number of possible frame alignments. Rather than matching individual frames of a test utterance with individual frames of the reference models and deriving scores for each matched pair of frames, this allows all frames of the test utterance to be compared and scored against every frame of the reference model, giving a statistical distribution of the frequency of occurrence of frame score values. "Good" frame matches will yield low scores and "poor" frame matches will yield high scores (or the converse, depending on the scoring scheme). A test utterance frame tested against a large number of reference models will result in a normal distribution as illustrated in Fig. 10. Most frame scores will lie close to the mean and within a few standard deviations therefrom. Because of the massive overlapping of frames in the reference models, the score distributions will include "best matches" between accurately aligned corresponding frames of the test utterance and reference models. When a test utterance from a particular speaker is tested against the reference model for that speaker, the distribution will thus include a higher incidence of very low scores. This ultimately results in "true speaker" scores being consistently low due to some

parts of the utterance being easily identified as originating from the true speaker while other parts less obviously from the true speaker are classified by being from the general population. Impostor frame scores will not produce low scores and will be classified as being from the general population.

That is, in accordance with this aspect of the invention, the reference models comprise sets of coefficients derived for a plurality of massively overlapping frames, and a test utterance is tested by comparing all of the frames of the test utterance with all of the frames of the relevant reference models and analysing the distribution of frame scores obtained therefrom.

The massive overlapping of frames applied to speech samples for enrolment purposes may also be applied to input utterances during subsequent speaker recognition, but this is not necessary.

The use of massive overlaps in the enrolment sample data is also beneficial in dealing with problems arising from noise present in periods of silence in the sample data. Such problems are particularly significant for text-independent speaker recognition systems. The existence of silences may or may not cause problems for an individual model or verification attempt, but they will cause deterioration in the overall system performance. The question is therefore how do we remove this completely or minimise the adverse effect. The use of massive frame overlaps in

the present invention contains an inherent solution. Consider the equations, which describe averaging the frame spectra (discussed in more detail below),

$$\overline{s(\omega)} = \frac{1}{N}\sum_n s_n(\omega) \quad = \quad \frac{1}{N}\sum_n (ss_n(\omega) \times sd_n(\omega))$$

$$= \frac{1}{N}(ss_1(\omega) \times sd_1(\omega)) + (ss_2(\omega) \times sd_2(\omega)) + ...(ss_N(\omega) \times sd_N(\omega))$$

$$= ss(\omega) \times \frac{1}{N}(sd_1(\omega) + sd_2(\omega) + ...sd_2(\omega)..... + sd_N(\omega))$$

It can be seen that the static parts (ss) average to $ss(\omega)$ and that individual frames have the spectra $ss_n(\omega) \times sd_n(\omega)$. Consider however the spectra of two added frames,

$(ss_1(\omega) \times sd_1(\omega)) + (ss_2(\omega) \times sd_2(\omega)) = ss(\omega) \times (sd_1(\omega) + sd_2(\omega))$

we have the steady part multiplied by a new spectra $sd_1(\omega) + sd_2(\omega)$. But since it is to be reduced by averaging, and it is also dynamic or variable in nature, the new spectra should behave in exactly the same way as a randomly extracted frame. The implication of this is that frames could be randomly added together with minimal effect on performance. This observation is not entirely true since we can have the case of valid speech frames added to silence frames in which the net result is a valid speech frame. This in fact results in an improvement in performance, as we are no longer including unwanted silences in the modelling.

If a typical signal with some minor silence problems has time frames randomly added, the silences would be eliminated but the signal would appear to have undergone major corruption. However the present invention using massively overlapped frames still functions. Interestingly, the implication of this is that channel echoes have no effect and can be ignored. It also underlines the fact that the preferred operating modes of the present invention extract the static parts of the spectra to a larger extent than conventional verifiers (as discussed further below). The addition of frames in this way has substantially the same effect as adding coloured noise to prevent unwanted modelling as discussed above.

In accordance with another aspect, the invention uses clustering or averaging techniques such as Vector Quantisation applied by the modules 28, 28a-n in a manner that differs from statistical analysis techniques used in conventional speaker recognition systems.

Preferably, the system of the present invention uses a Vector Quantisation (VQ) technique in processing the enrolment sample data output from the spectral analysis modules 26, 26a-n. This is a simplified technique, compared with statistical analysis techniques such as HMM employed in many prior art systems, resulting in transparent modelling providing models in a form which allow model-against-model comparisons in the subsequent

classification stage. Also, VQ as deployed in the present invention does not use temporal information, making the system resistant to impostors.

The VQ process effectively compresses the LPCC output data by identifying clusters of data points, determining average values for each cluster, and discarding data which do not clearly belong to any cluster. This results in a set of second matrices of second coefficients, representing the LPCC data of the set of first matrices, but of reduced size (typically, for example, 64 x 24 as compared with 1000 x 24).

The effects of the use of LPCC spectral analysis and clustering/averaging in the present invention will now be discussed.

The basic model assumes that spectral magnitude is useful and that the phase may be disregarded. This is known to apply to human hearing and if it was not applied to a verifier the system would exhibit undesirable phase related problems, such as sensitivity to the distance of the microphone from the speaker. Further assume that the spectral information of a speech sample can be regarded as consisting of two parts a static part $ss(\omega)$ and a dynamic part $sd(\omega)$ and that the processes are multiplicative. It is also assumed that the dynamic part is significantly larger than the static part.

$$s(\omega) = ss(\omega) \times sd(\omega)$$

As, by definition, the static part is fixed it is the more useful as a biometric as it will be related to the static characteristics of the vocal tract. This will relate the measure to some fixed physical characteristic as opposed to $sd(\omega)$ which is related to the dynamics of the speech.

The complete extraction of $ss(\omega)$ would give a biometric which exhibits the properties of a physical biometric, i.e. cannot be changed at will and does not deteriorate over time. Alternatively the exclusive use of $sd(\omega)$ will give a biometric which exhibits the properties of a behavioral biometric, i.e. can be changed at will and will deteriorate over time. A mixture of the two should exhibit intermediate properties but as $sd(\omega)$ is much larger than $ss(\omega)$ it is more likely that a combination will exhibit the properties of $sd(\omega)$; i.e. behavioral.

As with all frequency representations of a signal the assumption is that the time signal exists from $-\infty$ to $+\infty$ which clearly is not physically possible. In practice all spectral estimates of a signal will be made using a window, which exists for a finite period of time. The window can either be rectangular or shaped by a function (such as a Hamming window).

The use of a rectangular window amounts to simply taking a section of a signal in the area of interest and assuming that it is zero elsewhere. This

technique is common in speech processing in which the sections of signal are called frames; Fig. 1 shows a time signal with the frames indicated.

The frames can be shaped using an alternate window. Interestingly, the major effect of windowing is a spreading of the characteristic of a particular frequency to its neighbours, a kind of spectral averaging. This effect is caused by the main lobe; in addition to this the side lobes produce spectral oscillations, which are periodic in the spectrum. The present system later extracts the all-pole Linear Prediction coefficients, which have the intended effect of spectral smoothing and the extra smoothing, caused by the windowing, is not seen as a major issue. However, the periodic side lobe effects might be troublesome if the window size was inadvertently changed. This however can be avoided by good housekeeping.

Given that we can divide the time signal into frames the spectral characteristics for frames 1 to N can be represented as

$$s_1(\omega) = ss_1(\omega) \times sd_1(\omega) \quad ; \quad s_2(\omega) = ss_2(\omega) \times sd_2(\omega) \quad ; \quad \bullet \quad \bullet \quad \bullet$$

$$\bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet \quad \bullet$$

$$s_n(\omega) = ss_n(\omega) \times sd_n(\omega) \quad ; \quad \bullet \quad \bullet \quad \bullet \quad \bullet$$

$$s_N(\omega) = ss_N(\omega) \times sd_N(\omega)$$

But by definition

$$ss(\omega) = ss_1(\omega) = ss_2(\omega) = ss_3(\omega) \quad \bullet \quad \bullet \quad \bullet \quad = ss_N(\omega)$$

On first impressions to extract $ss(\omega)$ would seem to be possible using an averaging process,

$$\overline{s(\omega)} = \frac{1}{N}\sum_n s_n(\omega) \quad = \quad \frac{1}{N}\sum_n (ss_n(\omega) \times sd_n(\omega))$$

$$=$$

$$\frac{1}{N}(ss_1(\omega) \times sd_1(\omega)) + (ss_2(\omega) \times sd_2(\omega)) + ...(ss_N(\omega) \times sd_N(\omega))$$

$$=$$

$$ss(\omega) \times \frac{1}{N}(sd_1(\omega) + sd_2(\omega) + ...sd_2(\omega)..... + sd_N(\omega)) \quad = \quad ss(\omega) \times U(\omega)$$

where,

$$U(\omega) = \frac{1}{N}(sd_1(\omega) + sd_2(\omega) + ...sd_2(\omega)..... + sd_N(\omega))$$

If the frames had independent spectral characteristics (each resulting from random process) then $U(\omega)$ would tend to white noise, i.e. would have a flat spectrum so that $\overline{s(\omega)}$ could be extracted by smoothing the spectrum. This would most likely be the case if N were very large $\rightarrow \infty$. Given the linear nature of the *time domain - frequency domain - time domain* transformations a similar analysis could have been described in the time domain.

For real world conditions it cannot be assumed that N would be large in the sense that the frames have independent spectral characteristics. It is important to remember that this would require N to be large under two conditions:

1. During model creation
2. During a verification event

Failure to comply during either would potentially cause a system failure (error), however a failure in 1 is the more serious as it would remain a potential source of error until updated, whereas a problem in 2 is a single instance event.

If U(ω) cannot be guaranteed to converge to white noise, what can be done to cope with the situation? First consider that:

1. U(ω) will be a variable quantity
2. When smoothed across the frequency spectrum it would ideally be flat; i.e. the smoothed version $\mathrm{Usm}(\omega)=1$
3. U(ω) is the truncated sum of the speech frames the number of which would ideally tend to infinity.

Considering the equation

$$\overline{s(\omega)} = \mathrm{ss}(\omega) \times \frac{1}{N} \sum_n \mathrm{sd}_n(\omega)$$

The summation part tending to a flat spectrum is not an ideal performance measure, if we return to the frame based equivalent:

$$\overline{s(\omega)} = \frac{1}{N} \sum_n (\mathrm{ss}_n(\omega) \times \mathrm{sd}_n(\omega))$$

If we take the logarithms of the frames:

$$\frac{1}{N} \sum_n \log((ss_n(\omega) \times sd_n(\omega))$$

$$= \frac{1}{N} \sum_n \left[\log(\mathrm{ss}_n(\omega)) + \log(\mathrm{sd}_n(\omega))\right] = \log(\mathrm{ss}(\omega)) + \frac{1}{N} \sum_n \log(\mathrm{sd}_n(\omega))$$

$$= \mathrm{lss}(\omega) + \mathrm{lsd}(\omega)$$

it can be seen that the relationship between the static and dynamic parts is now additive. Because the relationship between the time domain and the frequency domain is linear a transformation from frequency to time gives:

$$\mathrm{lss}(\omega) + \mathrm{lsd}(\omega) \rightarrow \mathrm{cs}(\tau) + \mathrm{cd}(\tau) = \mathrm{c}(\tau)$$

In signal processing $c(\tau)$ is known as the Cepstral transformation of s(t) as discussed previously.

In general cepstral analysis consists of

$$\mathrm{time\_domain} \rightarrow \mathrm{frequency\_domain} \rightarrow \log(\mathrm{spectrum}) \rightarrow \mathrm{time\_domain}$$

The Cepstral transformation has been used in speech analysis in many forms.

As discussed above, in our current usage we create the Cepstral coefficients for the frames and extract the static part,

$$\frac{1}{N}\sum_n c_n(\tau) = \frac{1}{N}\sum_n (cs_n(\tau) + cd_n(\tau)) = cs(\tau) + \frac{1}{N}\sum_n cd_n(\tau)$$

Ideally the length of the speech signal would be long enough so that the dynamic part was completely random and the mean would tend to zero. This would leave the static part cs(t) as our biometric measure. However, we have a number of problems to overcome.

1. How do we handle the imperfect nature of the sum-to-zero

2. channel variation
3. endpointing
4. additive noise

Referring to the imperfect nature of the sum-to-zero, the nature of the Cepstral coefficients are such that they decay with increasing time and have the appearance of an impulse response for stable systems. This means that the dynamic range of each coefficient is different and they are in general in descending order.

It can be shown that the differences between the average coefficients of a test sample and the frame coefficient values for the true speaker model and the frame coefficient values of an impostor model are not large and a simple summation over all of the utterance frames to produce a distance score will be difficult to threshold in the conventional manner.

If we consider the two difficult problems associated with this methodology together rather than separately the answer to the problem is revealed. To re-emphasise, the two points of difficulty are,

1. the utterances will never be long enough for the mean of the dynamic part to converge to zero
2. the differences between the true speaker and the impostors will be small and difficult to threshold.

Consider two speakers with models based upon

$$\overline{c(\tau)} = \frac{1}{N}\sum_n c_n(\tau) = \frac{1}{N}\sum_n (cs_n(\tau) + cd_n(\tau)) = cs(\tau) + \frac{1}{N}\sum_n cd_n(\tau)$$

so that the models are m1(τ) and m2(τ), where,

$$m1(\tau) = \overline{c1(\tau)} = \frac{1}{N}\sum_n c1_n(\tau) = \frac{1}{N}\sum_n (csl_n(\tau) + cdl_n(\tau)) = csl(\tau) + \frac{1}{N}\sum_n cdl_n(\tau)$$

$$= csl(\tau) + el(\tau) \quad ; \text{ where } e1(\tau) \text{ is the error}$$

In vector form the models are

$$m1 = \begin{bmatrix} csl_1 + el_1 \\ csl_2 + el_2 \\ \bullet \\ \bullet \\ csl_p + el_p \end{bmatrix} \quad \text{and} \quad m2 = \begin{bmatrix} cs2_1 + e2_1 \\ cs2_2 + e2_2 \\ \bullet \\ \bullet \\ cs2_p + e2_p \end{bmatrix}$$

A test utterance from speaker 1 expressed in the same form will be

$$T1 = \begin{bmatrix} csl_1 + Tel_1 \\ csl_2 + Tel_2 \\ \bullet \\ \bullet \\ csl_p + Tel_p \end{bmatrix}$$

using a simple distance measure true speaker distance is

$$d1 = |m1 - T1| = \left\| \begin{bmatrix} csl_1 + el_1 \\ csl_2 + el_2 \\ \bullet \\ \bullet \\ csl_p + el_p \end{bmatrix} - \begin{bmatrix} csl_1 + Tel_1 \\ csl_2 + Tel_2 \\ \bullet \\ \bullet \\ csl_p + Tel_p \end{bmatrix} \right\| = |el - Tel|$$

impostor distance is

$$d2 = |m2 - Tl| = \left\| \begin{bmatrix} cs2_1 + e2_1 \\ cs2_2 + e2_2 \\ \bullet \\ \bullet \\ cs2_p + e2_p \end{bmatrix} - \begin{bmatrix} csl_1 + Tel_1 \\ csl_2 + Tel_2 \\ \bullet \\ \bullet \\ csl_p + Tel_p \end{bmatrix} \right\| = |cs2 - csl + e2 - Tel|$$

Assuming that the convergence of the dynamic parts of the models is good (i.e. that the error vectors are small compared to the static vectors) then in general d1<d2. This is simply stating that the models built represent the enrolled speaker (a condition that can easily checked during enrolment using the data available at that time). Interestingly, if e1 and e2 are small compared to the test signal error Tel the distances become independent of e1 and e2. The condition under which the test error will be large when compared to the model error is during text-independent test conditions. This shows that if the dynamic components of the enrolment speech samples are minimised in the enrolment models then such models can provide a good basis for text-independent speaker recognition

The errors e1 and e2 above are average model construction errors; the actual errors are on a frame by frame basis and will have a distribution about the mean. This distribution could be modelled in a number of ways the simplest being by use of a standard clustering technique such as k-means to model the distribution. The use of k-means

clustering is also known in other forms as Vector Quantisation (VQ) and is a major part of the Self Organising Map (SOM) also known as the Kohonen Artificial Neural Network.

The system just described where a test utterance is applied to two models and the closest chosen is a variant of identification. In the above case if either speaker 1 or speaker 2, the enrolled speakers, claim to be themselves and are tested they will always test as true and so the False Rejection Rate FRR =0. If an unknown speaker claims to be either speaker1 or speaker2 he will be classified as one or the other, so there is a 1/2 chance of success and hence a False Acceptance Rate FAR =50%. If an equal number of true speaker tests and random impostor tests were carried out, we can calculate an overall error rate as (FRR+FAR)/2 = (0+0.5)/2=25%

It is obvious that the number of models (the cohort) against which the test utterance is tested will have an effect on the FAR and it will reduce as the cohort increases. It can be shown that the accuracy of recognition under these conditions is asymptotic to 100% with increasing cohort size, since FRR=0, but as the accuracy is

$$\text{accuracy} = 100 - (\text{FRR} + \text{FAR})\frac{100}{2} = 100 - \left(\text{FRR} + \frac{1}{\text{cohort\_size}}\right)\frac{100}{2}$$

it is in more general terms asymptotic to 100-FRR.

It is worth observing at this point that the FRR and FAR are largely decoupled: the FRR is fixed by the quality of the model produced and the FAR is fixed by the cohort size. It is also worth observing that to halve the error rate we need to double the cohort size e.g. for 99% accuracy the cohort is 50, for 99.5% accuracy the cohort is 100, for 99.75% accuracy the cohort is 200. As the cohort increases the computational load increases and in fact doubles for each halving of the error rate. As the cohort increases to very large numbers the decoupling of the FRR and FAR will break down and the FRR will begin to increase.

Rather than continually increasing the cohort size in an attempt to reduce the FAR to a minimum another approach is needed. The approach, in accordance with one aspect of the invention, is to use parallel processes (also discussed elsewhere in the present description), which exhibit slightly different impostor characteristics and are thus partially statistically independent with respect to the identifier strategy. The idea is to take a core identifier which exhibits the zero or approximately zero FRR and which has a FAR that is set by the cohort size. The front end processing of this core identifier is then modified slightly to reorder the distances of the cohort member models from the true speaker model. This is done while maintaining the FRR~0 and can be achieved by altering the spectral shaping filters 24a-24n (see Fig. 7), or by altering

the transformed coefficients, such as by using delta-ceps etc.

When an enrolled speaker uses the system the test signal is applied to all of the processes in parallel but each process has a FRR~0 and the speaker will pass. When an unknown impostor uses the system he will pass each individual process with a probability of 1/cohort_size. However with the parallel processes we have introduced conditional probabilities. That is, if an impostor passes process1 what is the likelihood of him passing the modified process2 as well etc. Although the probability of an impostor passing all of the processes is not that of the statistically independent case of

$$\text{statistically\_independent\_result} = \text{process\_prob}^{\text{no\_of\_processes}}$$

it does however reduce with the addition of processes. It can be shown that for a given process FAR value, the overall accuracy of the system increases with the number of processes.

Where multiple parallel processes are used in this way, the scheme for matching a test sample against a claimed identity may require a successful match for each process or may require a predetermined proportion of successful matches.

The combined use of massive sample frame overlaps and Vector Quantisation (or equivalent) in building enrolment models in accordance with the present invention provides particular advantages. The

massive overlapping is applied at the time of constructing the models, although it could also be applied at the time of testing an utterance. The technique involves using a massive frame overlap, typically 80-90%, to generate a large possible number of alignments; the frames generated by the alignments are then transformed into representative coefficients using the LPCC transformation to produce a matrix of coefficients representing all of the alignments. This avoids conventional problems of frame alignment. The matrix is typically of the size no_of_frames by LPCC_order, for example 1000x24. This is repeated for all of the utterances used in constructing the model, typically 10, Giving 10 matrices of 1000x24. Vector Quantisation is then used to reduce the data to produce a model for the speaker. This has the effect of averaging the frames so as to reduce the significance of the dynamic components of the sampled speech data as discussed above. The resulting model does not take cognisance of the frame position in the test utterance and is hence not temporal in nature. This addresses the problem of temporal dependency.

The combined use of VQ and massive frame overlapping produces an operation mode which is different from conventional systems based upon HMM/DTW. In HMM/DTW all frames are considered to be equally valid and are used to form a final score for thresholding into a yes/no decision. In the present invention every row (frame) of the test sample data is tested against every row of the enrolment model data for

the claimed speaker and the associated impostor cohort. For each row of the test sample data, a best match can be found with one row of the enrolment model, yielding a test score for the test sample against each of the relevant enrolment models. The test sample is matched to the enrolment model that gives the best score. If the match is with the claimed identity, the test speaker is accepted. If the match is with an impostor the test speaker is rejected.

The present system, then, uses LPCC and VQ modelling (or similar/equivalent spectral analysis and clustering techniques), together with massive overlapping of the sample frames, to produce the reference models for each enrolled speaker, which are stored in the database. In use of the system, an input test utterance is subjected to similar spectral analysis to obtain an input test model which can be tested against the enrolled speaker data-set. Advantageously, this approach can be applied so as to obtain a very low False Rejection Rate (FRR), substantially equal to zero. The significance of this is discussed further below.

Parallel Modelling

As previously discussed, the performance of speaker recognition systems in accordance with the invention can be improved by using multiple parallel processes to generate the model.

Referring now to Fig. 7 of the drawings, one preferred embodiment of a speaker recognition system employing parallel modelling processes in accordance with one aspect of the invention comprises an input channel 100 for inputting a signal representing a speech sample to the system, a channel normalisation process 200 as described elsewhere, a plurality of parallel signal processing channels 102a, 102b ... 102n, a classification module 110 and an output channel 112. The system further includes an enrolled speaker data-set 114; i.e. a database of speech models obtained from speakers enrolled to use the system. The speech sample data is processed in parallel by each of the processing channels 102a-n, the outputs from each of the processing channels is input to the classification module 110, which communicates with the database 114 of enrolled speaker data, and a decision as to the identity of the source of the test utterance is output via the output channel 112.

Each of the processing channels 102a-n comprises, in series, a spectral shaping filter 24a-n, an (optional) added noise input 206a-n, as described elsewhere, a spectral analysis module 26a-n and a statistical analysis module 28a-n. The outputs from each of the statistical analysis modules 28a-n is input to the classification module 110.

The spectral shaping filters 24a-n comprise a bank of filters which together divide the utterance signal into a plurality of overlapping frequency

bands, each of which is then processed in parallel by the subsequent modules 26a-n and 28a-n. The number of processing channels, and hence the number of frequency bands, may vary, with more channels providing more detail in the subsequent analysis of the input data. Preferably, at least two channels are employed, more preferably at least four channels. The filters 24a-n preferably constitute a low-pass or band-pass or high-pass filter bank. The bandwidth of the base filter 24a is selected such that the False Rejection Rate (FRR) resulting from subsequent analysis of the output from the first channel 102a is zero or as close as possible to zero. The subsequent filters 24b-n have incrementally increasing bandwidths that incrementally pass more of the signal from the input channel 100. The FRR for the output from each channel 102a-n is thus maintained close to zero whilst the different channel outputs have slightly different False Acceptance (FA) characteristics. Analysis of the combined outputs from the channels 102a-n yields a reduced overall FA rate (a claimed identity is only accepted if the outputs from all of the channels are accepted) with a FRR close to zero. The significance of this is discussed further below.

The use of multiple frequency bands improves upon conventional single-channel spectral analysis, increasing the size of the feature vectors of interest in the subsequent statistical analysis.

It will be understood that different types of parallel processing may be employed in the modelling process in order to provide multiple feature sets modelling different (related or unrelated) aspects of the input speech sample and/or alternative models of similar aspects. Banks of filters of other types in addition to or instead of low pass filters might be employed. Different types or variants of spectral and/or statistical analysis techniques might be used in parallel processing channels. Parallel statistical analyses may involve applying different weighting values to sets of feature coefficients so as to obtain a set of slightly deviated models.

It will be understood that the architecture illustrated in Fig. 7 may be used for both obtaining enrolment models for storing in the database 114 and for processing test speech samples for testing against the enrolment models. Each enrolment model may include data-sets for each of a plurality of enrolment utterances. For each enrolment utterance, there will be a matrix of data representing the output of each of the parallel modelling processes. Each of these matrices represents the clustered/averaged spectral feature vectors. Test sample data is subject to the same parallel spectral analysis processes, but without clustering/averaging, so that the test model data comprises a matrix representing the spectral analysis data for each of the parallel modelling processes. When a test model is tested against an

enrolment model, the test matrix representing a particular modelling process is tested against enrolment matrices generated by the same modelling process.

CLASSIFICATION

The nature of the reference models obtained by the modelling techniques described above is such that they lend themselves to direct model against model comparisons. This enables the system to employ an identifier strategy in which each enrolment model is associated with an impostor cohort. That is, for the reference model of each enrolled speaker ("subject"), there is an impostor cohort comprising a predetermined number of reference models of other enrolled speakers, specific to that subject and which has a known and predictable relationship to the subject's reference model. These predictable relationships enable the performance of the system to be improved. Fig. 11(a) shows the results obtained by a conventional speaker recognition system, similar to Fig. 3, comparing scores for an input utterance tested against reference data for eight speakers. Speaker 1 is the true speaker, but the scores for some of the other speakers are sufficiently close to reduce significantly the degree of confidence that the system has identified the correct speaker. Fig. 11(b) shows equivalent results obtained using a system in accordance with the present invention. It can be seen that the results for speaker 1 are much more clearly

distinguished from the results of all of the other speakers 2 to 8.

The speaker modelling method employed in the preferred embodiments of the present invention is inherently simpler (and, in strict mathematical terms, cruder) than conventional techniques such as HMM and possible alternatives such as gaussian mixture models. However, the present applicants believe that the conventional use of "tight" statistical methods is inherently flawed and result in poor "real world" performance, and that, surprisingly, the relatively simpler statistical methods of the present invention are much more effective in practice. As previously noted, the temporal nature of HMM makes it susceptible to mimics, a problem which is avoided by the present invention. Further, the models of the present invention are ideally suited to enable analysis of the structure of the enrolled speaker data-set by model against model testing.

The ability to perform model against model comparisons by using the present speaker models provides two particular advantages. Firstly, this provides the ability to identify the most relevant impostors in the enrolled speaker data-set (i.e. those which are close to and uniformly distributed around a particular model) and to produce an effective and predictable speaker normalisation mechanism. VQ modelling involves choosing the size of the model; i.e. choosing the number of

coefficients ("centres"). Once this has been done, the positions of the centres can be moved around until they give the best fit to all of the enrolment data vectors. This effectively means allocating a centre to a cluster of enrolment vectors, so each centre in the model represents a cluster of information important to the speaker identity.

The model against model tests make it possible to predict how an enrolled speaker, or claimed identity, will perform against the database both in the broad sense and in an area local (in the system dataspace) to the claimed identity. Fig. 12 illustrates the results of testing reference models for speakers 2 to 8 against the reference model for speaker 1. The ellipses show the model against model results whilst the stars show actual scores for speaker utterances tested against model 1. It can be seen that the model against model tests can be used to predict the actual performance of a particular speaker against a particular reference model. The model against model results tend to lie at the bottom of the actual score distributions and therefore indicate how well a particular impostor will perform against model 1. This basic approach of using model against model tests to predict actual performance is known as such. As described further below, this approach may be extended in accordance with one aspect of the present invention to guard particular models against impostors using individually selected, statistically variable groupings.

The second advantage derived from model against model testing is the ability to predict the performance of a test utterance against some or, if need be, all of the enrolled speaker models. This enables a virtually unlimited number of test patterns to be used to confirm an identity, which is not possible with conventional systems.

In addition, the model against model test results may be used to assemble a specific impostor cohort for use with each reference model. This allows accurate score normalisation and also allows each model to be effectively "guarded" against impostors by using a statistically variable grouping which is selected for each enrolled speaker. This is illustrated by Fig. 13. Each reference model can be regarded as a point in a multi-dimensional dataspace, so that "distances" between models can be calculated. Fig. 13 illustrates this idea in two dimensions for clarity, where each star represents a model and the two-dimensional distance represents the distance between models.

It can be seen that the distribution of speaker models is not uniform, so that a world-model based normalisation technique will not operate equally well for all speaker models. It can also be seen that some speaker models can be relatively close to one another, which implies that there is potential for impostors to successfully impersonate enrolled speakers. For each speaker model, these issues can

be resolved by creating a specific cohort of impostors around the subject model. This simplifies normalisation and creates a guard against impostors. This is illustrated in Fig. 14, which shows, in a similar manner to Fig. 13, a subject model represented by a circle, members of an impostor cohort represented by stars, and a score for an impostor claiming to be the subject, represented by an "x". The impostor score is sufficiently close to the subject model to cause recognition problems. However, because the speaker data-set enables prediction of how the true subject speaker will perform against the models of the impostor cohort, this information can be used to distinguish the impostor x from the true subject, by testing the impostor against the models of the cohort members as well as against the true subject model. That is, it can be seen that the impostor utterance x is closer to some of the cohort members than would be expected for the true subject, and further away from others than expected. This would indicate an impostor event and result in the impostor utterance being rejected as a match for the true subject.

This provides the basis for a two stage recognition process which firstly rejects impostors who are clearly not the claimed speaker followed, where necessary, by a more detailed process applied to utterances which are close enough to possibly be the claimed speaker.

In certain applications of speaker verification systems, it is important to minimise the possibility of "false rejections"; i.e. instances in which the identity claimed by a user is incorrectly rejected as being false. In accordance with one aspect of the invention, an "identifier strategy" is employed which provides very low false rejections, whilst also providing predictable system performance and minimising problems associated with the use of thresholds in accepting or rejecting a claimed identity.

In accordance with this strategy, the database of enrolled speakers (the "speaker space") is partitioned; e.g. so that each speaker enrolled in the system is assigned to a cohort comprising a fixed number N of enrolled speakers, as described above. The speaker classification module of the system (e.g. the module 110 in the system of Fig. 4) operates such that the input test utterance is compared with all of the members of the cohort associated with the identity claimed by the speaker, and the test utterance is classified as corresponding to that member of the cohort which provides the best match. That is, the test utterance is always matched to one member of the cohort, and will never be deemed not to match any member of the cohort. If the cohort member to which the utterance is matched corresponds to the claimed identity, then the claimed identity is accepted as true. If the utterance is matched to any other

member of the cohort then the claimed identity is rejected as false.

The modelling and classification processes can be tuned such that the proportion of false rejections is effectively zero (FR = 0%) (as discussed above); i.e. there is substantially zero probability that a speaker will be wrongly identified as a member of the cohort other than the claimed identity. This is facilitated by the use of model against model comparisons such that a match is not based simply upon the test utterance being matched against the single closest model, but also on the basis of its relationship to other members of the cohort. Where the cohort is of a fixed size N, the maximum possible proportion of false acceptances FA = 100/N % and the total average error rate = (FA + FR)/2 = 50/N %. If the cohort size N is 20, the error rate is thus 2.5 %; i.e. an accuracy of 97.5 %. If the cohort size is fixed, the system is scalable to any size of population while maintaining a fixed and predictable error rate. That is, the accuracy of the system is based on the size of the cohort and is independent of the size of the general population, making the system scalable to very large populations. Accuracy can be improved by increasing the cohort size, as long as the false rejection rate does not increase significantly.

This strategy does not rely on the use of thresholds to determine a result, but thresholds could still be used to reduce false acceptances; i.e. once a test

utterance has been matched to the claimed identity using the foregoing strategy, thresholds could be applied to determine whether the match is close enough to be finally accepted.

As indicated above, the selection of an impostor cohort associated with a particular enrolment model may involve the use of algorithms so that the members of the impostor cohort have a particular relationship with the enrolment model in question. In principle, this may provide a degree of optimisation in the classification process. However, it has been found that a randomly selected impostor cohort performs equally well for most practical purposes. The most important point is that the cohort size should be predetermined in order to give predictable performance. The impostor cohort for a particular enrolment model may be selected at the time of enrolment or at the time of testing a test utterance.

Parallel Classification

The performance of a speaker recognition system in accordance with the invention may be improved by the use of multiple parallel classification processes. Generally speaking, such processes will be statistically independent or partially independent. This approach will provide multiple classification results which can be combined to derive a final result, as illustrated in Fig. 5.

In one example, using the identifier strategy described above, the same test utterance may be tested against a number of different cohorts, or against different enrolment phrases, or combinations thereof. Where multiple cohorts are employed, each cohort will give a result with a false rejection rate of essentially zero (FR = 0 %) and a false acceptance rate FA = 100/N % as before. The overall false acceptance rate for n cohorts of equal size will be
FA = $100*M/N^n$ % and the average error rate
= $50*M/N^n$ %, where M is a coefficient having a value greater than 1 and representing the effect of the processes not being entirely statistically independent. That is, with 2 cohorts and a cohort size of 20, the average error rate will be 0.125*M % as compared with 2.5 % for a single cohort as described above. Thresholds may also be applied to further improve accuracy as previously described.

Other types of partially statistically independent processes may be employed in the modelling process, the classification process or both as previously discussed. Besides the examples previously given, a single utterance may be divided into parts and processed separately.

NORMALISATION

A further problem encountered with conventional speaker recognition systems is that system performance may be affected by differences between

speech sampling systems used for initial enrolment and subsequent recognition. Such differences arise from different transducers (microphones), soundcards etc. In accordance with a further aspect of the present invention, these difficulties can be obviated or mitigated by normalising speech samples on the basis of a normalisation characteristic which is obtained and stored for each sampling system (or, possibly, each type of sampling system) used to input speech samples to the recognition system. Alternatively (preferably), the normalisation characteristic can be estimated "on the fly" when a speech sample is being input to the system. The normalisation characteristic(s) can then be applied to all input speech samples, so that reference models and test scores are independent of the characteristics of particular sampling systems. Alternatively or additionally, in accordance with a further aspect of the invention a normalisation process can be applied at the time of testing test sample data against enrolment sample data.

A normalisation characteristic is effectively a transfer function of the sampling system and can be derived, for example, by inputting a known reference signal to the sampling system, and processing the sampled reference signal through the speech recognition system. The resulting output from the recognition system can then be stored and used to normalise speech samples subsequently input through the same sampling system or the same type of sampling system.

Alternatively, as illustrated in Fig. 15, a speech signal S(f) which has been modified by the transfer function C(f) of an input channel 300 can be normalised on the fly by inputting the modified speech signal S(f)*C(f) to an estimating module 302, which estimates the transfer function C(f) of the channel 300, and to a normalisation module 304, and applying the inverse of the estimated transfer function 1/C(f) to the normalisation module, so that the output from the normalisation module closely approximates the input signal S(f). The estimator module 302 creates a digital filter with the spectral characteristics of the channel 300 and the inverse of this filter is used to normalise the signal. For example, the inverse filter can be calculated by determining the all-pole filter which represents the spectral quality of a sample frame. The filter coefficients are then smoothed over the frames to remove as much of the signal as possible, leaving the spectrum of the channel (C(f)). The estimate of the channel spectrum is then used to produce the inverse filter 1/C(f). This basic approach can be enhanced to smooth the positions of the poles of the filters obtained for the frames, with intelligent cancellation of the poles to remove those which are known not to be concerned with the channel characteristics.

Depending on the nature of the transfer function/normalisation characteristic, the normalisation process can be applied to the speech

sample prior to processing by the speaker recognition system or to the spectral data or to the model generated by the system.

A preferred method of channel normalisation, in accordance with one aspect of the invention, is applied to the test model data and the relevant enrolment models at the time of testing the test sample against the enrolment models.

The overall effect of the channel characteristics on a speech signal could be described as

$$\hat{s}(\omega) = ss(\omega) \times sd(\omega) \times cc(\omega)$$

where $\hat{s}(\omega)$ is the estimate of the speakers characteristics, $cc(\omega)$ is the channel characteristic or changed channel characteristic as appropriate, and the speech signal is treated as comprising a static part and a dynamic part as before. Ideally the unwanted channel characteristic can be estimated and removed. In practice the removal can be achieved in the time domain, frequency domain or a combination. They both achieve the same effect, that is to estimate $cc(\omega)$ and remove it using some form of inverse filter or spectral division. If $\hat{cc}(\omega)$ is the estimate of the spectrum of the unwanted channel then we would calculate

$$\frac{\hat{s}(\omega)}{\hat{cc}(\omega)} = ss(\omega) \times sd(\omega) \times \frac{cc(\omega)}{\hat{cc}(\omega)} \approx s(\omega)$$

If the estimation of the channel characteristic is good $\frac{cc(\omega)}{\hat{c}c(\omega)} \approx 1$ and our estimate of the speech is good with the unwanted spectral shaping removed. This would normally be implemented using a algorithm based on the FFT.

An alternative implementation is to model the channel characteristic as a filter, most likely in the all-pole form,

$$h(z) = \frac{z^N}{z^N + a_{N-1}z^{N-1} + ...a_0}$$

This is the most basic form of the ARMA and would normally be extracted from the time signal directly, possibly using Linear Prediction.

A similar normalisation could be carried out on the Cepstral representation.

In the Cepstral domain the speech signal is represented as

$$c(\tau) = cs(\tau) + cd(\tau)$$

and the speech signal modified by the unwanted channel characteristics is

$$\hat{c}(\tau) = cs(\tau) + cd(\tau) + cc(\tau)$$

It can be seen that in this case we have an additive process rather than a product. But it should also be remembered that both cs and cc are static and we may need to remove one *cc* without removing the other.

It is important to consider the context in which we would wish to remove the signal cc and their different conditions (enrolled model, database derived cohort, test speaker etc.).

Figure 16 illustrates various sources of corruption of a speech sample in a speaker recognition system. The input speech signal s(t) is altered by environmental background noise, b(t), the recording device bandwidth, r(t), electrical noise and channel crosstalk, t(t), and transmission channel bandwidth, c(t), so that the signal input to the recognition system is an altered signal v(t). The system is easier to analyse in the frequency domain and the signal at the verifier is:

$$v(\omega) = ((s(\omega) + b(\omega)).r(\omega) + t(\omega)).c(\omega) \qquad \text{eq1}$$

At the verifier we can define two conditions, when the person is speaking and when he is not. Resulting in two equations,

$$v(\omega) = ((s(\omega) + b(\omega)).r(\omega) + t(\omega)).c(\omega)$$

and

$$v(\omega) = ((0 + b(\omega)).r(\omega) + t(\omega)).c(\omega)$$

First consider the simplified problem as it applies to the systems in accordance with the present invention; assume that b(t)=t(t)=0

$$v(\omega) = s(\omega).r(\omega).c(\omega) = s(\omega).h(\omega)$$

where h( ) is the combined channel spectral characteristic,

$$h(\omega) = r(\omega).c(\omega)$$

$$v(\omega) = s(\omega).h(\omega) = ss(\omega).sd(\omega).h(\omega)$$

The cohort models are selected from the database of speakers recorded using the same channel (b) and the true speaker model is recorded using a different channel (a). The test speaker can either be the true speaker or an impostor and will be recorded using a third channel (c). Figure 17 shows this diagrammatically. Fig. 18 shows the same thing expressed in the alternate form using the Cepstral coefficients. It should be remembered that the values of the signal components as represented in Figs 17 and 18 are averages corresponding to the summations of sample frame data.

Consider the claimed identity model, which was built from,

$$v_1(\tau) = cs_1(\tau) + cd_1(\tau) + h_a(\tau) \qquad \text{eq2}$$

and the cohort models which were built from,

$$v_m(\tau) = cs_m(\tau) + cd_m(\tau) + h_b(\tau) \qquad \text{eq3}$$

The problem for the verifier is that there are two different channels used in the identifier and if we assume the difference between them is

$$hd(\tau) = h_a(\tau) - h_b(\tau)$$

or

$$h_a(\tau) = h_b(\tau) + hd(\tau)$$

then the claimed identity model referred to the cohort channel (b) will be

$$v_1(\tau) = cs_1(\tau) + cd_1(\tau) + h_a(\tau) = cs_1(\tau) + cd_1(\tau) + h_b(\tau) + hd(\tau)$$

and

$$v_1(\tau) = (cs_1(\tau) + hd(\tau)) + cd_1(\tau) + h_b(\tau)$$

it can be seen that the mean of the static part of the claimed identity model has been shifted by the difference between the channels and will cause an error if the true speaker is tested using channel-b if the situation is not corrected. Similar problems involving false acceptances using channel-a will also occur.

One method of addressing this problem is to remove the mean from the claimed identity model, but a simple removal of the mean would at first glance produce,

$$v_t(\tau) = cd_t(\tau)$$

where the static part of the speaker model has also been removed. However, examining equation 1 (the system model including additive noise)

$$v(\omega) = ((s(\omega) + b(\omega)).r(\omega) + t(\omega)).c(\omega)$$

if we consider the case during which the speaker pauses, $s(\omega)=0$

then $$v(\omega) = (b(\omega).r(\omega) + t(\omega)).c(\omega)$$

and $$v(\omega) = n(\omega).c(\omega)$$

where $n(\omega)$ is a noise signal.

In cepstral form this would be

$$v(\tau) = n(\tau) + c(\tau) = sn(\tau) + dn(\tau) + c(\tau)$$

where as before sn is the static part of the noise and dn is the result of the summation of the dynamic part.

The average of a model constructed from this would be

$$sn(\tau) + c(\tau)$$

where sn is any steady state noise such as an interference tone and c is the channel.

Considering again equation1 (the claimed identity model build conditions)

$$v_1(\tau) = cs_1(\tau) + cd_1(\tau) + h_a(\tau)$$

this was the noise free case, adding a steady state noise gives,

$$v_1(\tau) = cs_1(\tau) + cd_1(\tau) + h_a(\tau) + sn(\tau)$$

If we constructed the speaker pause model for this case we would get

$$sn(\tau) + h_a(\tau)$$

and using this to remove the mean results in

$$v_1(\tau) = cs_1(\tau) + cd_1(\tau)$$

This gives us a model unbiased by the channel. A similar process could be applied to each model whereby it has the channel bias removed by its own silence model. The test speaker could be similarly treated, i.e. its silence model is used to remove the channel effects.

The removal (reduction) of the channel characteristics using the silence model as described above requires suitable channel noise and perfect detection of the silence parts of the utterance. As these cannot be guaranteed they need to be mitigated (for instance, if the silence includes some speech we will include some of the claimed identity speaker

static speech and inadvertently remove it). Fortunately they can be dealt with in one simple modification to the process: the cohort models should all be referred to the same silence model.

That is, if we re-add the silence average of the claimed identity model to all of the models in the cohort (including the claimed identity model). This refers all of the models to the same mean $sn(\tau)+h_a(\tau)$. This normalisation is also applied to the test model, thereby referring all of the models and the test utterance to the same reference point. In effect we choose a reference channel and noise condition and refer all others to it.

This is illustrated diagrammatically in Fig. 19, which shows the Cepstral coefficients of the test utterance together with the claimed identity model and the cohort models 1 to m being input to the classifier 110. A "silence model" or "normalisation model" 400 derived from the claimed identity enrolment data is used to normalise each of these before input to the classifier, so that the actual inputs to the classifier are a normalised test utterance, normalised claimed identity model and normalised cohort models. Ideally, the normalisation model 400 is based on data from periods of silence in the claimed identity enrolment sample as discussed above, but it could be derived from the complete claimed identity enrolment sample. In practical terms, the normalisation model comprises a single row of Cepstral coefficients,

each of which is the mean value of one column (or selected members of one column) of Cepstral coefficients from the claimed identity model. These mean values are used to replace the mean values of each of the sets of input data. That is, taking the test utterance as an example, the mean value of each column of the test utterance Cepstral coefficients is subtracted from each individual member of that column and the corresponding mean value from the normalisation model is added to each individual member of the column. A similar operation is applied to the claimed identity model and each of the cohort models.

It will be understood that the normalisation model could be derived from the claimed identity model or from the test utterance or from any of the cohort models. It is preferable for the model to be derived from either the claimed identity model or the test utterance, and it is most preferable for it to be derived from the claimed identity model. The normalisation model could be derived from the "raw" enrolment sample Cepstral coefficients or from final model after Vector Quantisation. That is, it could be derived at the time of enrolment and stored along with the enrolment model or it could be calculated when needed as part of the verification process. Generally, it is preferred that a normalisation model is calculated for each enrolled speaker at the time of enrolment and stored as part of the enrolled speaker database.

These normalisation techniques can be employed with various types of speaker recognition systems but are advantageously combined with the speaker recognition systems of the present invention.

Speaker recognition systems in accordance with the invention provide improved real world performance for a number of reasons. Firstly, the modelling techniques employed significantly improve separation between true speakers and impostors. This improved modelling makes the system less sensitive to real world problems such as changes of sound system (voice sampling system) and changes of speaker characteristics (due to, for example, colds etc.). Secondly, the modelling technique is non-temporal in nature so that it is less susceptible to temporal voice changes, thereby providing longer persistence of speaker models. Thirdly, the use of filter pre-processing allows the models to be used for variable bandwidth conditions; e.g. models created using high fidelity sampling systems such as multimedia PCs will work with input received via reduced bandwidth input channels such as telephony systems.

It will be understood that the preferred methods in accordance with the present invention are inherently suited for use in text-independent speaker recognition systems as well as text-dependent systems.

SYSTEMS

The invention thus provides the basis for flexible, reliable and simple voice recognition systems operating on a local or wide area basis and employing a variety of communications/input channels. Fig. 16 illustrates one example of a wide area system operating over local networks and via the Internet, to authenticate users of a database system server 400, connected to a local network 402, such as an Ethernet network, and, via a router 404, to the Internet 406. A speaker authentication system server 408, implementing a speaker recognition system in accordance with the present invention, is connected to the local network for the purpose of authenticating users of the database 400. Users of the system may obviously be connected directly to the local network 402. More generally, users at sites such as 410 and 412 may access the system via desktop or laptop computers 414, 416 equipped with microphones and connected to other local networks which are in turn connected to the Internet 406. Other users such as 418, 420, 422 may access the system by dial-up modem connections via the public switched telephone network 424 and Internet Service Providers 426.

IMPLEMENTATION

The algorithms employed by speaker recognition systems in accordance with the invention may be implemented as computer programs using any suitable programming language such as C or C++, and executable programs may be in any required form

including stand alone applications on any hardware/operating system platform, embedded code in DSP chips etc. (hardware/firmware implementations), or be incorporated into operating systems (e.g. as MS Windows DLLs). User interfaces (for purposes of both system enrolment and subsequent system access) may similarly be implemented in a variety of forms, including Web based client server systems and Web browser-based interfaces, in which case speech sampling may be implemented using, for example, ActiveX/Java components or the like.

Apart from desktop and laptop computers, the system is applicable to other terminal devices including palmtop devices, WAP enabled mobile phones etc. via cabled and/or wireless data/telecommunications networks.

APPLICATIONS

Speaker recognition systems having the degree of flexibility and reliability provided by the present invention have numerous applications. One particular example, in accordance with a further aspect of the present invention, is in providing an audit trail of users accessing and/or modifying digital information such as documents or database records. Such transactions can be recorded, providing information regarding the date/time and identity of the user, as is well known in the art. However, conventional systems do not normally verify or authenticate the identity of the user.

Speaker recognition, preferably using a speaker recognition system in accordance with the present invention, may be used to verify the identity of a user whenever required; e.g. when opening and/or editing and/or saving a digital document, database record or the like. The document or record itself may be marked with data relating to the speaker verification procedure, or such data may be recorded in a separate audit trail, providing a verified record of access to and modification of the protected document, record etc. Unauthorised users identified by the system will be denied access or prevented from performing actions which are monitored by the system.

Improvements and modifications may be incorporated without departing from the scope of the invention as defined in the appended claims.

Claims

1. A method of processing speech samples to obtain a model of a speech sample for use in a speaker recognition system, comprising:

dividing the speech sample into a plurality of frames;

for each frame, obtaining a set of feature vectors representing the smoothed frequency spectrum of the frame;

applying a clustering algorithm to the feature vectors of the frames to obtain a reduced data set representing the original speech sample;

wherein the adjacent frames are overlapped by at least 80%.

2. The method of claim 1, wherein the adjacent frames are overlapped by less than 95%.

3. The method of claim 1, wherein the adjacent frames are overlapped by an amount in the range 80% to 90%.

4. The method of any preceding claim, wherein the clustering algorithm comprises a Vector Quantisation algorithm or a k-means algorithm.

5. The method of any preceding claim, wherein the set of feature vectors representing the smoothed frequency spectrum of the frame is obtained by means of Linear Predictive Coding/Cepstral analysis [LPCC]

or Fast Fourier Transform [FFT] or by use of a filter bank.

6. The method of any preceding claim, further including storing the model of the speech sample and the identity of the speaker in a database of enrolment models of speakers enrolled in a speaker recognition system.

7. The method of claim 6, wherein each enrolment model comprises a plurality of speech sample models representing a plurality of different utterances.

8. The method of claim 6 or claim 7, wherein each enrolment model comprises a plurality of speech sample models representing the same utterance modelled using a plurality of parallel modelling processes.

9. The method of any of claims 6 to 8, further including associating the model of the speech sample with a cohort comprising a predetermined number of other speakers enrolled in the speaker recognition system.

10. The method of any of claims 6 to 9, including processing a second speech sample to obtain a test model of the second speech sample for testing against said database of enrolment models, wherein processing said second speech sample comprises:

dividing the second speech sample into a plurality of frames; and

for each frame, obtaining a set of feature vectors representing the smoothed frequency spectrum of the frame.

11. The method of claim 10, wherein the set of feature vectors representing the smoothed frequency spectrum of the frame of the second sample is obtained by means of Linear Predictive Coding/Cepstral analysis [LPCC] or Fast Fourier Transform [FFT] or by use of a filter bank.

12. The method of claim 10 or claim 11, wherein the test model comprises a plurality of speech sample models representing the same utterance modelled using a plurality of parallel modelling processes.

13. The method of any of claims 10 to 12, wherein the identity of the speaker of the second speech sample is tested by testing the test model against the enrolment model for the claimed identity and the associated cohort as defined in claim 9.

14. A method of speaker recognition in which a plurality of speakers to be recognised by a speaker recognition system are enrolled by storing an enrolment model for each speaker in a database of enrolled speakers, the enrolment model representing at least one speech sample from that speaker, wherein:

each enrolled speaker is associated with a cohort of a predetermined number of other enrolled speakers;

a test speech sample from a speaker claiming to be one of the enrolled speakers is modelled and tested, using a classification process, against the enrolment model of the claimed speaker and the enrolment models of the associated cohort; and

the classification process always matches the test model with either the claimed speaker or one of the associated cohort such that a false acceptance rate of the system is determined by the cohort size.

15. The method of claim 14, wherein modelling processes used for modelling the enrolled speaker speech samples and the test speech sample and/or classification processes used for testing the test model against the enrolment models are selected to provide a false rejection rate substantially equal to zero, so that an overall error rate of the system is determined substantially only by the false acceptance rate.

16. The method of claim 14 or claim 15, wherein the test model is tested using multiple parallel classification processes and the test model is matched with an enrolment model only if at least a predetermined number of the parallel classification processes produces a match with that enrolment model, so as to reduce the false acceptance rate of the system for a given cohort size.

17. The method of claim 16, wherein the enrolment models and test model are each obtained using multiple parallel modelling processes and the

parallel classification processes compare the results of the parallel modelling processes applied to the test speech sample with corresponding results of the parallel modelling processes applied to the enrolment speech samples.

18. The method of any one of claims 8, 12 or 17, wherein the parallel modelling processes comprise at least one of:

different frequency banding applied to the speech samples;

different spectral modelling applied to the speech samples; and

different clustering applied to the feature vectors representing the speech samples.

19. The method of claim 16, wherein the parallel classification processes comprise testing the test model against different cohorts of enrolled speakers.

20. The method of claim 16, wherein the parallel classification processes comprise testing the test model against different utterances represented by the enrolment models.

21. The method of any of claims 14 to 20, wherein the enrolment models and test model are obtained using the method of any of claims 1 to 13.

22. A method of normalising speech models in a speaker recognition system of the type in which

speech samples are input to the system via different input channels having different channel characteristics, and wherein a test model representing a test sample is tested, using a classification process, against a set of enrolment models representing speech samples from speakers enrolled in the system, comprising deriving a normalisation model from the test speech sample or from one of the enrolment speech samples and using the normalisation model to normalise the test model and the enrolment models against which the test model is to be tested prior to testing the normalised test model against the normalised enrolment models.

23. The method of claim 22, wherein the normalisation model is derived from the enrolment speech sample for the identity claimed for the test speech sample.

24. The method of claim 23, wherein the normalisation model is derived from the enrolment model for the identity claimed for the test speech sample.

25. The method of any of claims 22 to 24, wherein the speech samples are divided into a plurality of frames, a set of feature vectors are obtained representing the smoothed frequency spectrum of each frame, and the normalisation model is obtained by calculating the mean values of sets of feature vectors from at least some of said frames of the

speech sample from which the normalisation model is derived.

26. The method of claim 25, wherein the frames used for deriving the normalisation model are frames corresponding to periods of silence in the speech sample from which the normalisation model is derived.

27. The method of claim 25 or claim 26, wherein the test model and enrolment models are normalised by replacing mean values of the feature vectors of the test model and the enrolment models with the corresponding mean values from the normalisation model.

28. The method of any of claims 22 to 27, wherein the speech samples are processed using the method of any of claims 1 to 13.

29. The method of any of claims 14 to 21, wherein the test model and enrolment models are normalised prior to classification using the method of any of claims 22 to 28.

30. The method of any preceding claim, wherein speech samples are input to a speaker recognition system via an input channel having a transfer function which modifies the speech sample data, comprising estimating the transfer function of said input channel and normalising the modified speech

sample data using the inverse of said estimated transfer function.

31. A speaker recognition system comprising data processing and storage means adapted to implement the method of any of claims 1 to 30.

32. A computer program comprising symbolic code for instructing a computer to execute the method of any of claims 1 to 30.

33. A data carrier encoded with a computer program comprising symbolic code for instructing a computer to execute the method of any of claims 1 to 30.

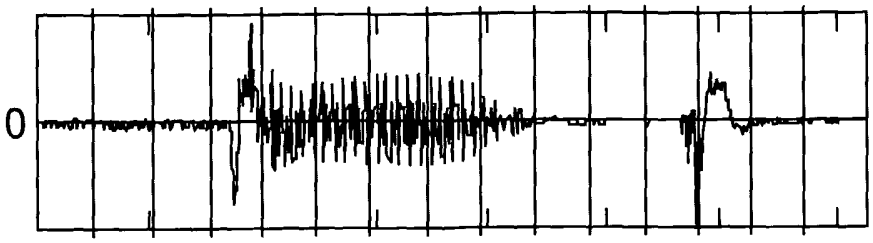

Fig. 1

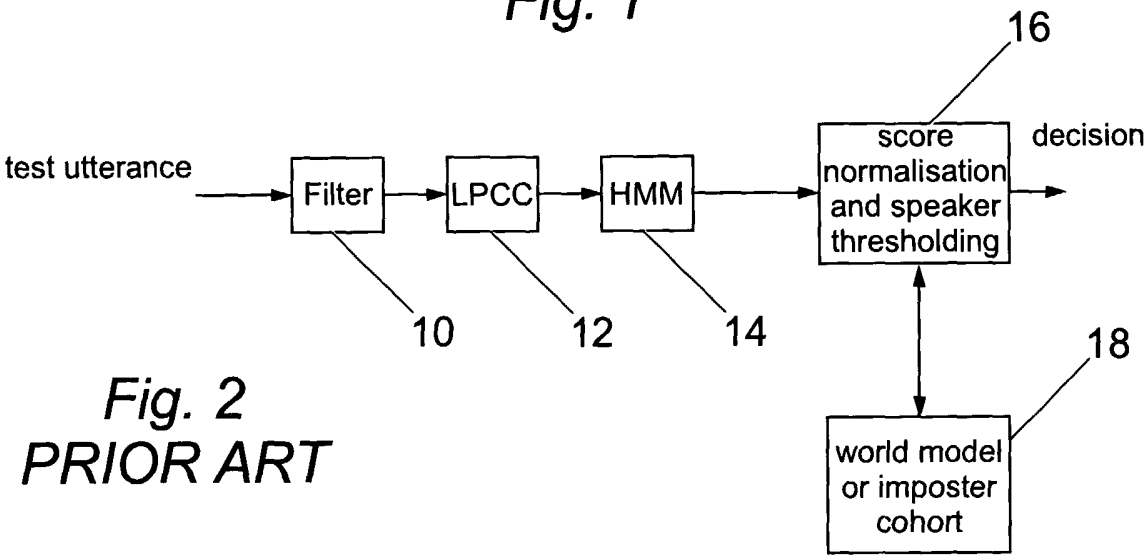

Fig. 2
PRIOR ART

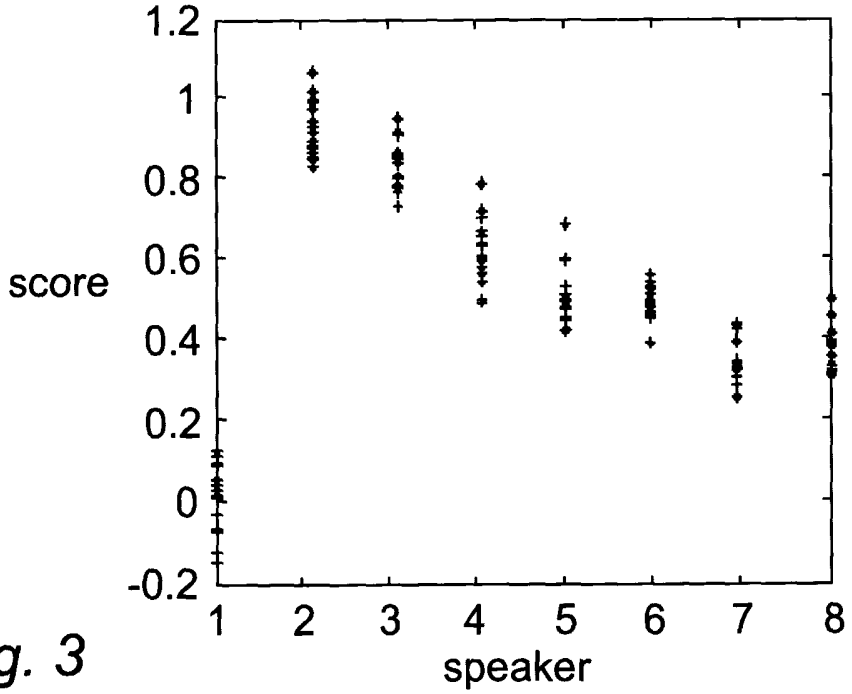

Fig. 3

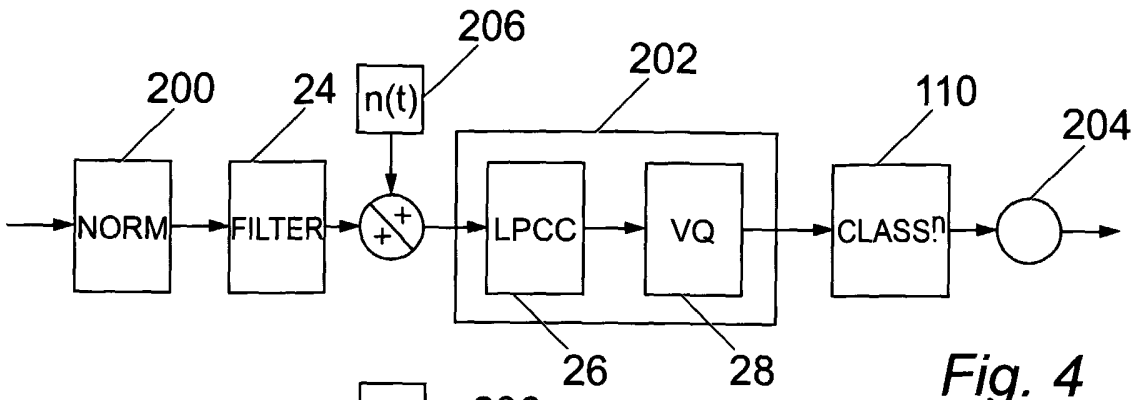

Fig. 4

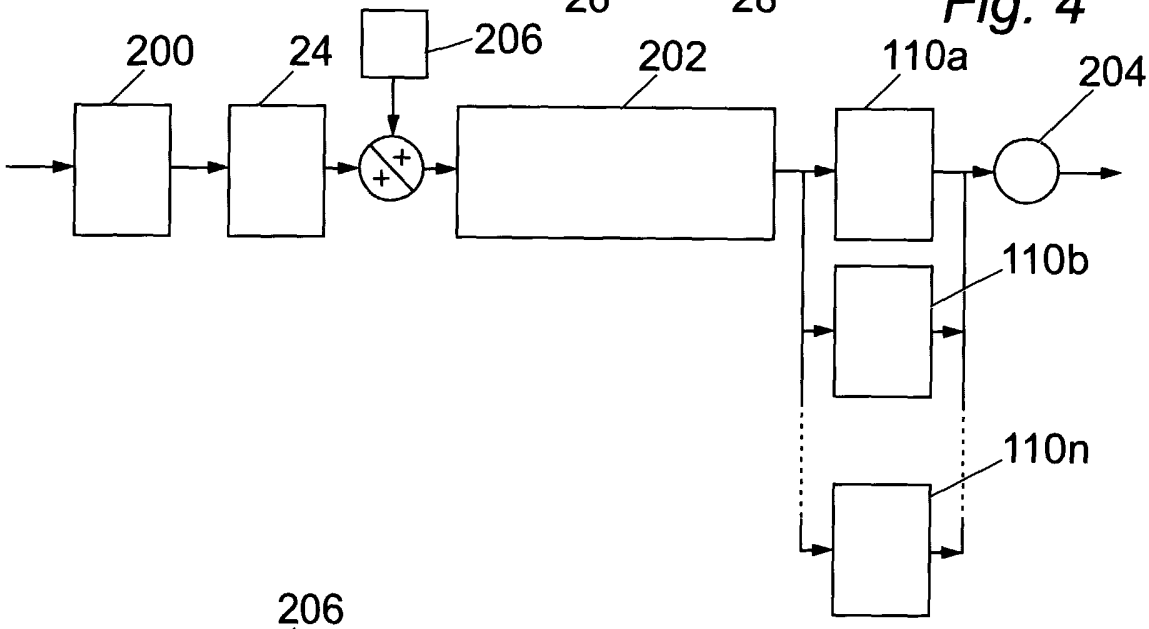

Fig. 5

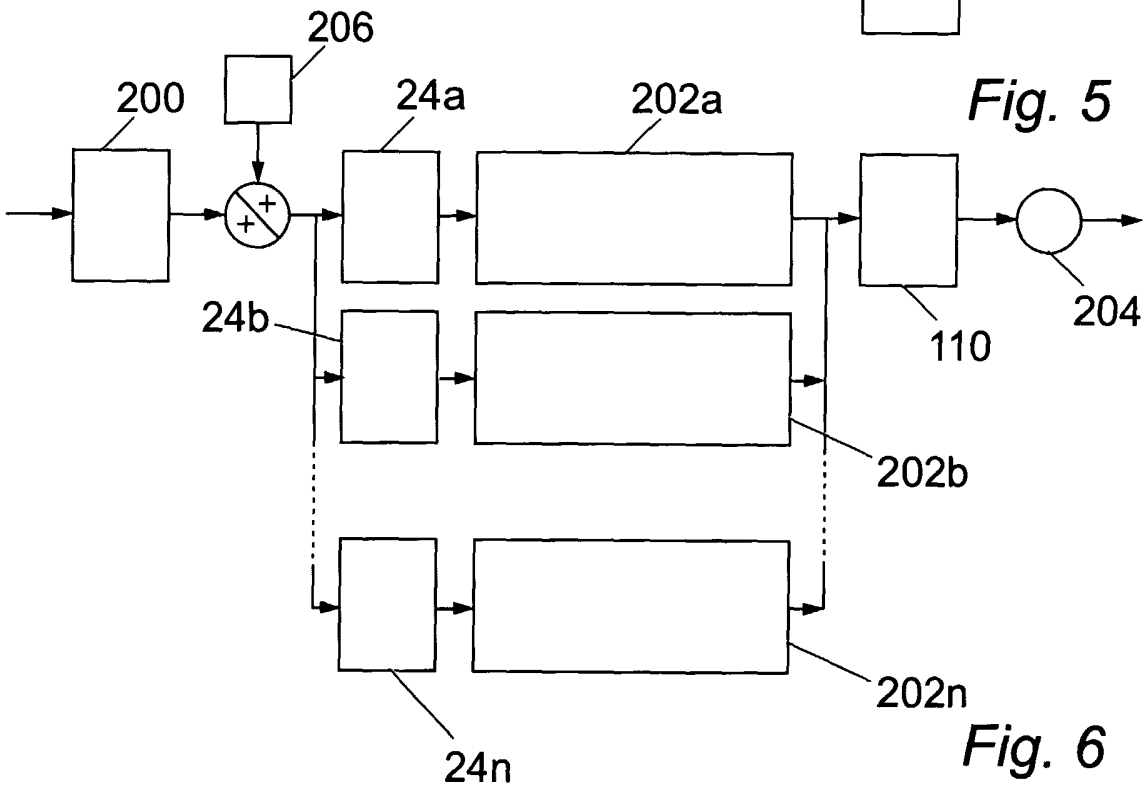

Fig. 6

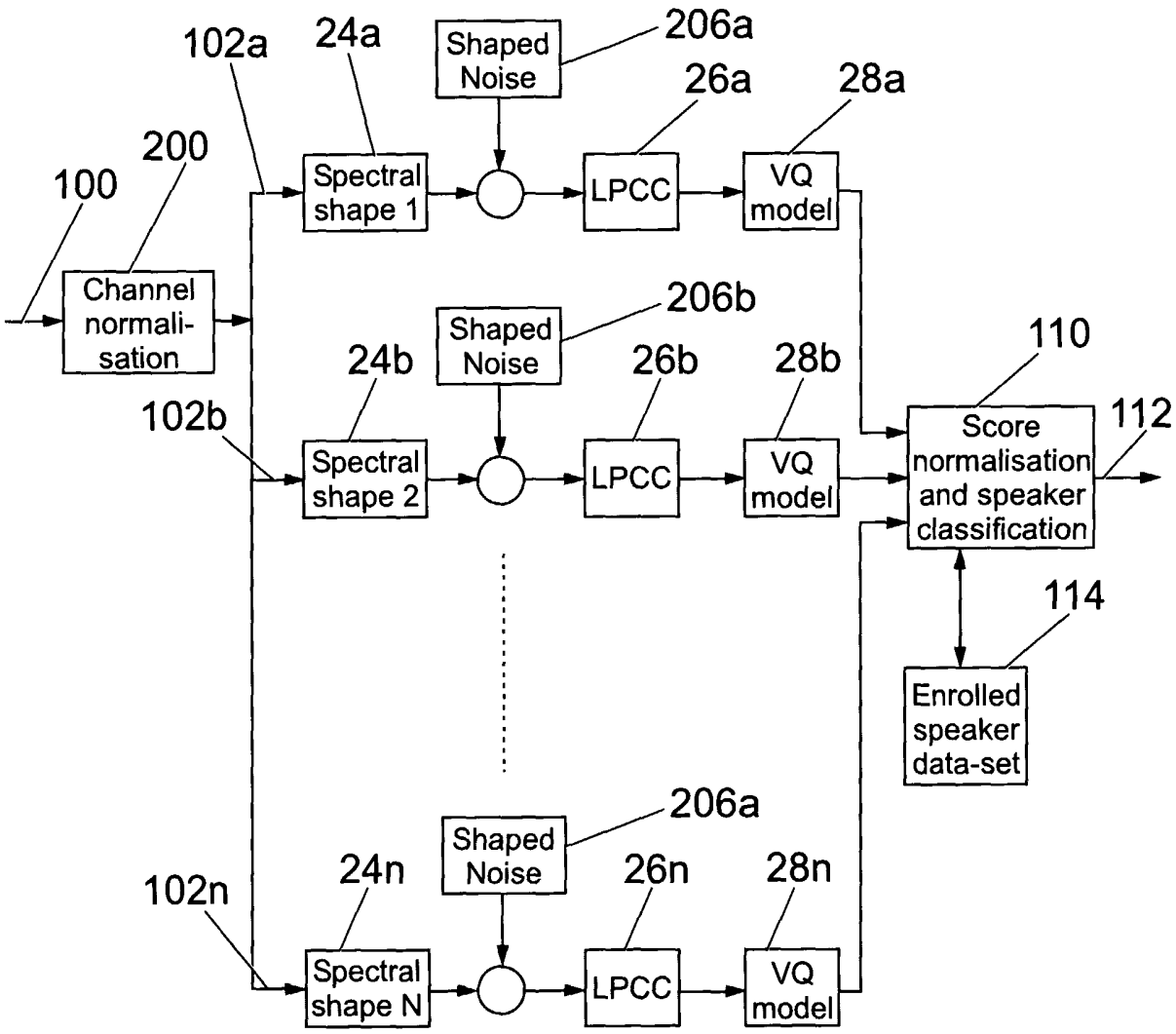
102a
24a
Shaped Noise
206a
26a
28a
200
100
Spectral shape 1
LPCC
VQ model
Channel normali-sation
206b
Shaped Noise
110
24b
26b
28b
102b
Score normalisation and speaker classification
112
Spectral shape 2
LPCC
VQ model
114
Enrolled speaker data-set
206a
Shaped Noise
24n
26n
28n
102n
Spectral shape N
LPCC
VQ model

Fig. 7

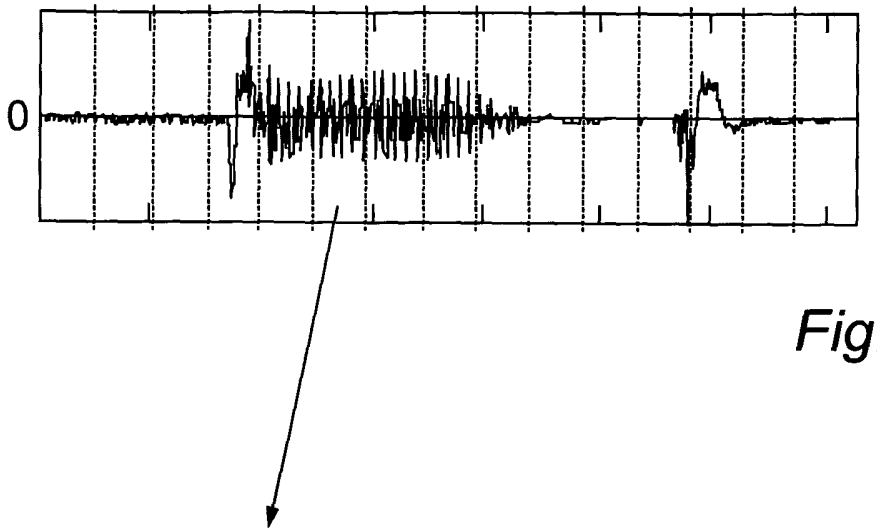

Fig. 8a

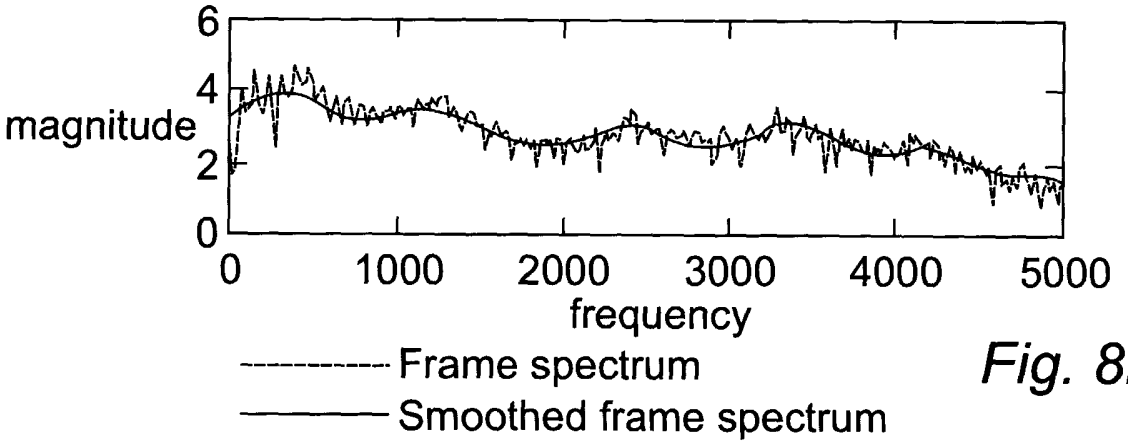

Fig. 8b

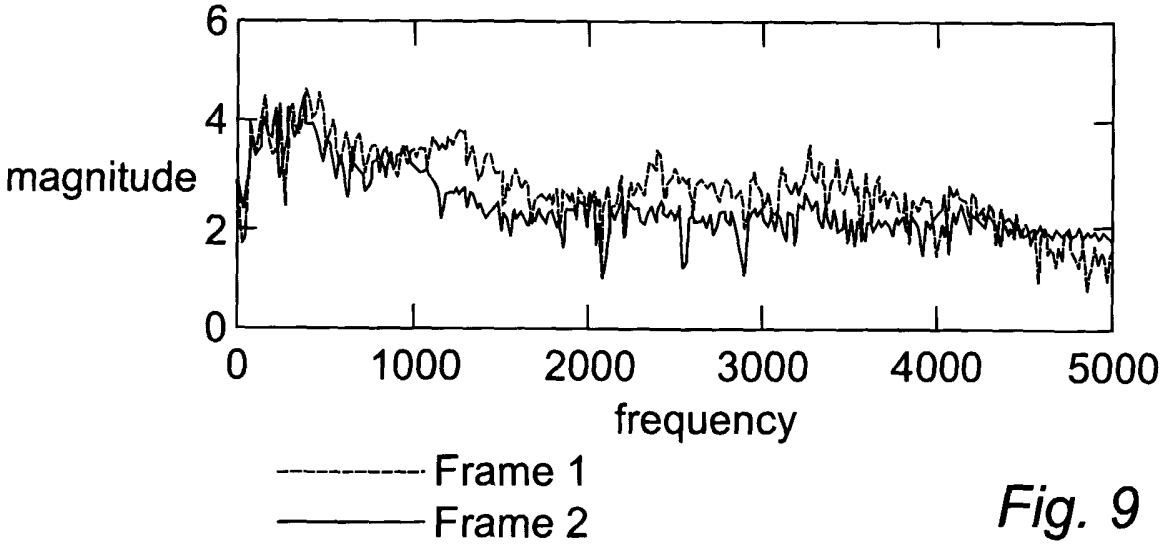

Fig. 9

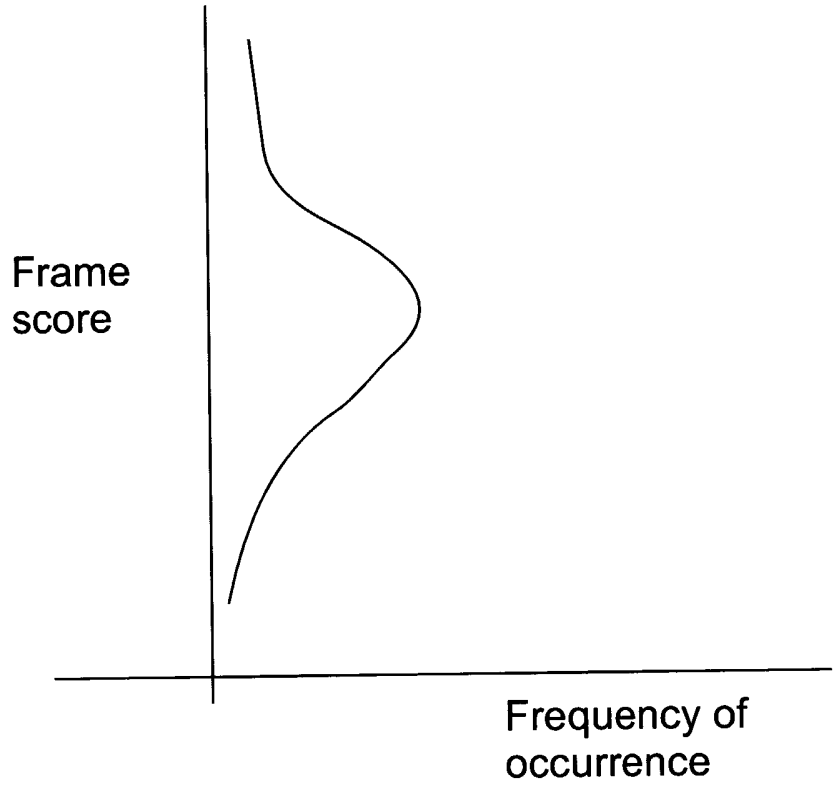

Fig. 10

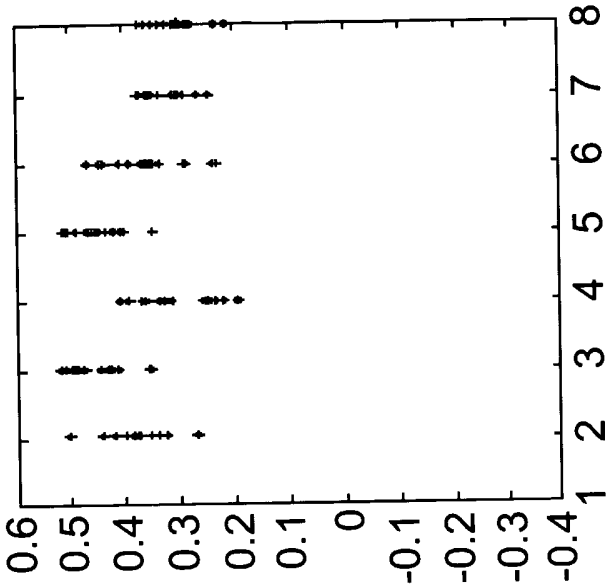

Fig. 11b

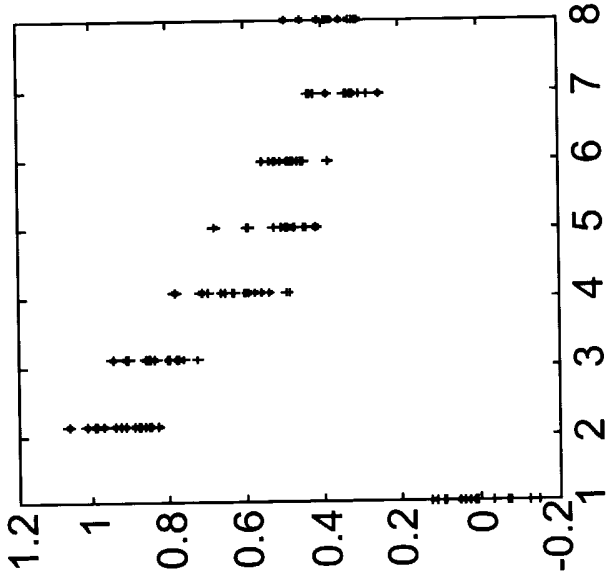

Fig. 11a

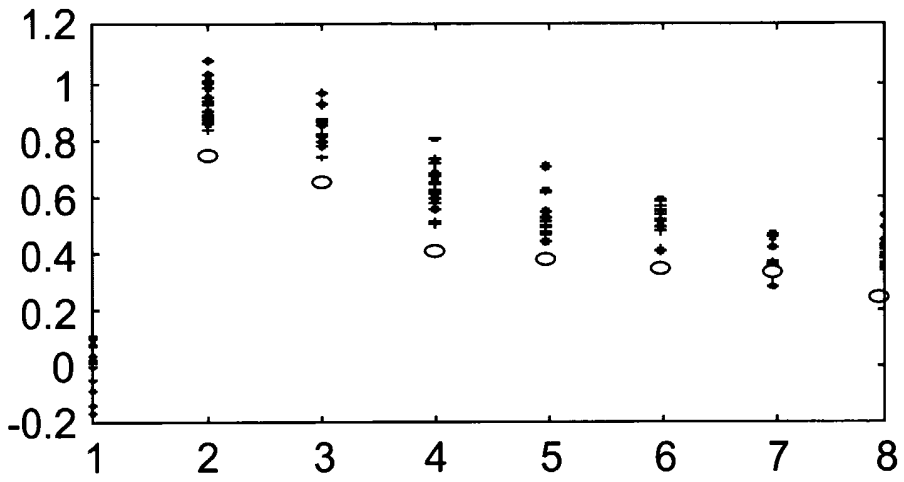

Fig. 12

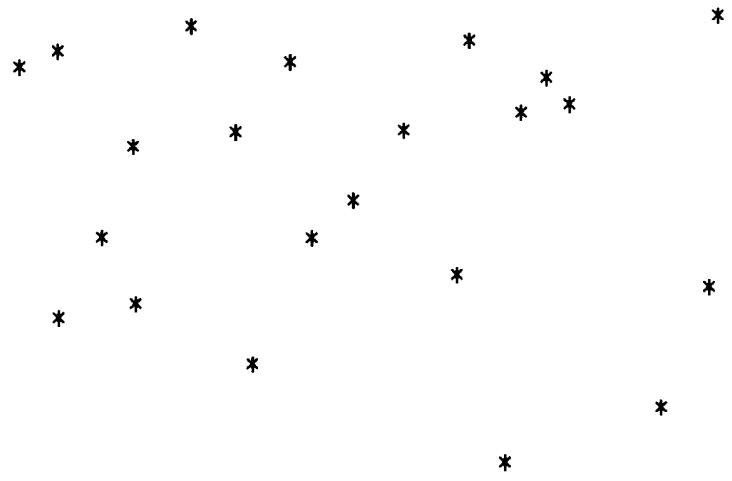

Fig. 13

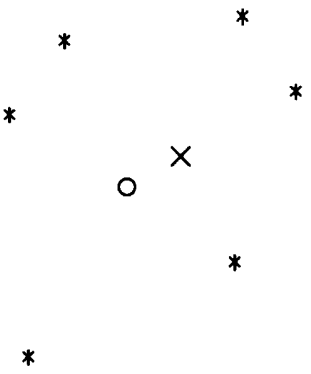

Fig. 14

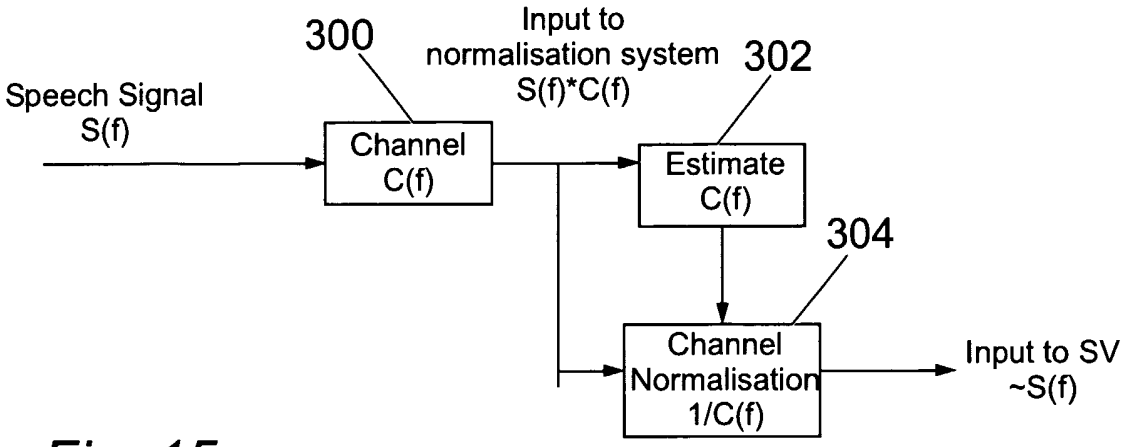
300
Input to
normalisation system
S(f)*C(f)
302
Speech Signal
S(f)
Channel
C(f)
Estimate
C(f)
304
Channel
Normalisation
1/C(f)
Input to SV
~S(f)

Fig. 15

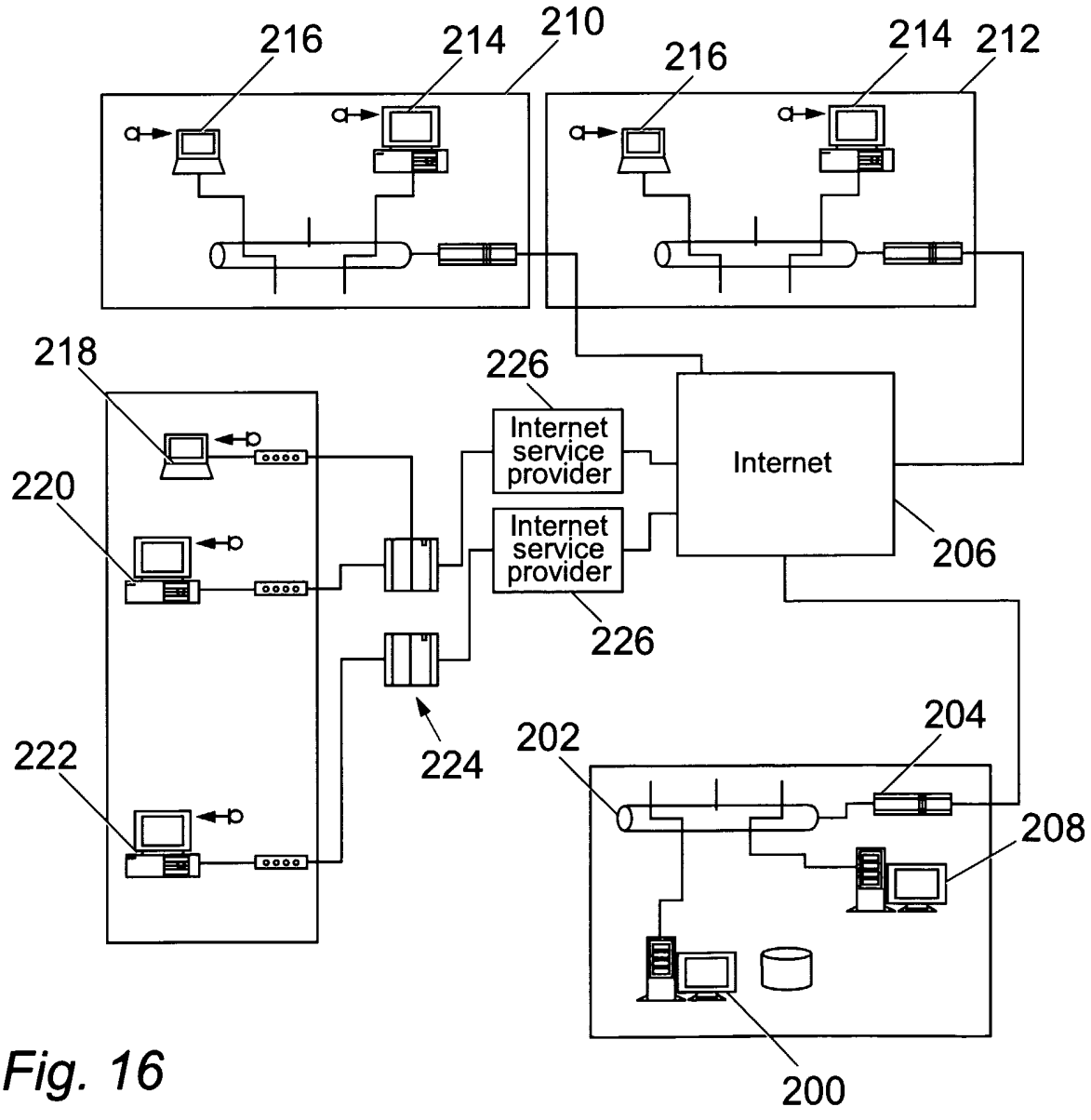
216
214
210
216
214
212
218
226
Internet
service
provider
Internet
220
Internet
service
provider
206
226
224
222
204
202
208
200

Fig. 16

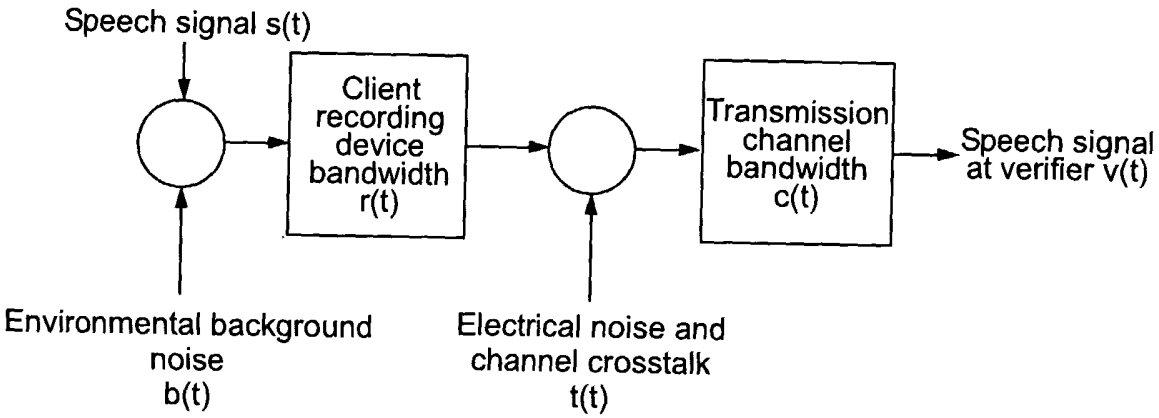
Speech signal s(t)
Client recording device bandwidth r(t)
Transmission channel bandwidth c(t)
Speech signal at verifier v(t)
Environmental background noise b(t)
Electrical noise and channel crosstalk t(t)

Fig. 17

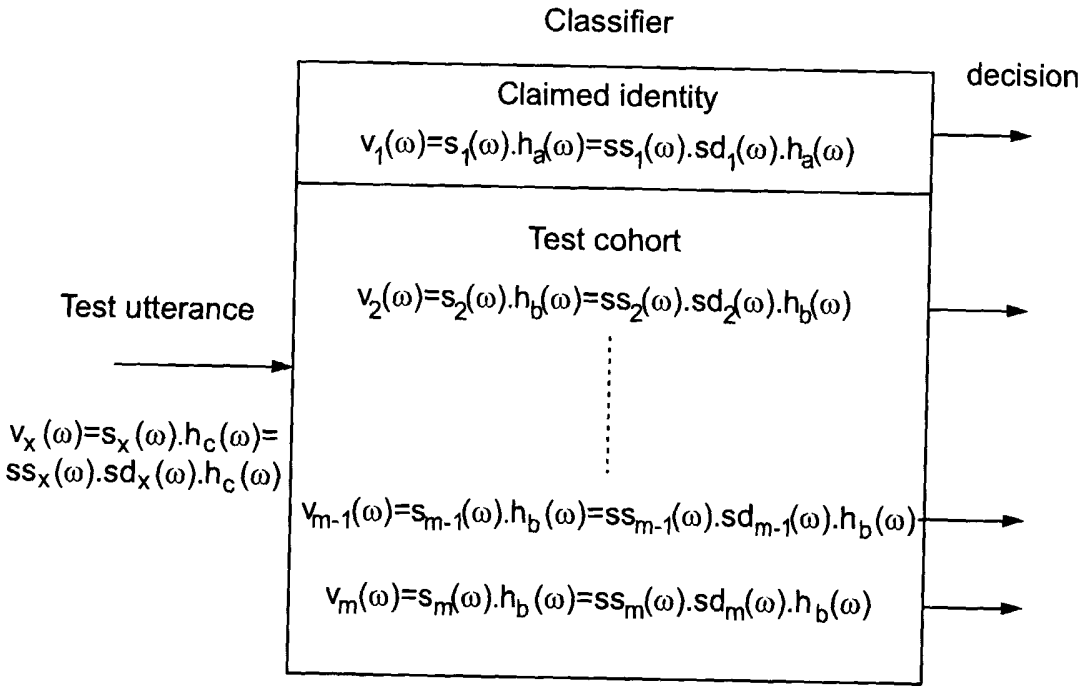
Classifier
Claimed identity
$v_1(\omega)=s_1(\omega).h_a(\omega)=ss_1(\omega).sd_1(\omega).h_a(\omega)$
Test cohort
$v_2(\omega)=s_2(\omega).h_b(\omega)=ss_2(\omega).sd_2(\omega).h_b(\omega)$
$v_{m-1}(\omega)=s_{m-1}(\omega).h_b(\omega)=ss_{m-1}(\omega).sd_{m-1}(\omega).h_b(\omega)$
$v_m(\omega)=s_m(\omega).h_b(\omega)=ss_m(\omega).sd_m(\omega).h_b(\omega)$
decision
Test utterance
$v_x(\omega)=s_x(\omega).h_c(\omega)=ss_x(\omega).sd_x(\omega).h_c(\omega)$

Fig. 18

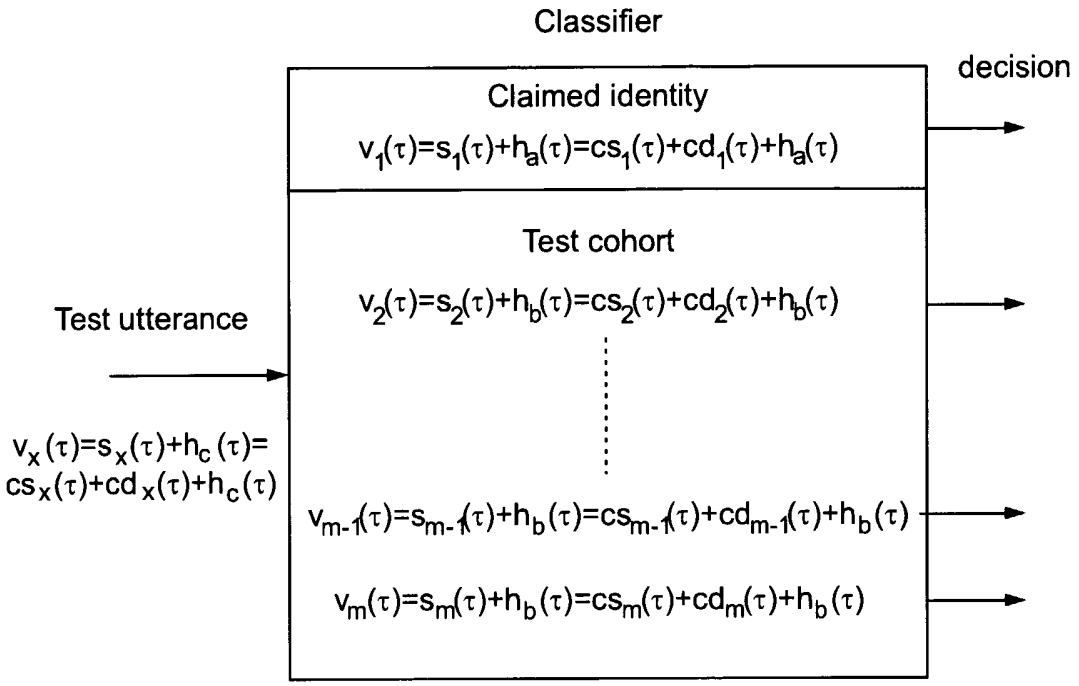
Classifier
decision
Claimed identity
$v_1(\tau)=s_1(\tau)+h_a(\tau)=cs_1(\tau)+cd_1(\tau)+h_a(\tau)$
Test cohort
$v_2(\tau)=s_2(\tau)+h_b(\tau)=cs_2(\tau)+cd_2(\tau)+h_b(\tau)$
Test utterance
$v_x(\tau)=s_x(\tau)+h_c(\tau)=$
$cs_x(\tau)+cd_x(\tau)+h_c(\tau)$
$v_{m-1}(\tau)=s_{m-1}(\tau)+h_b(\tau)=cs_{m-1}(\tau)+cd_{m-1}(\tau)+h_b(\tau)$
$v_m(\tau)=s_m(\tau)+h_b(\tau)=cs_m(\tau)+cd_m(\tau)+h_b(\tau)$

Fig. 19

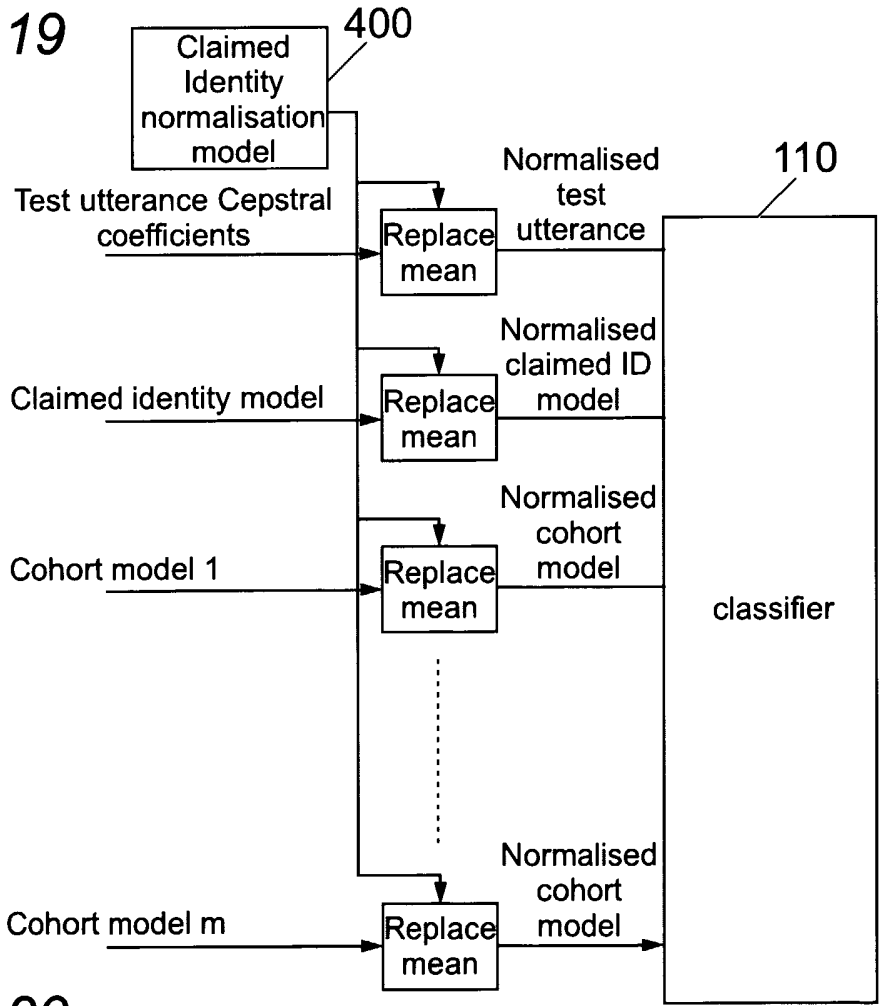
400
Claimed Identity normalisation model
110
Test utterance Cepstral coefficients
Replace mean
Normalised test utterance
Claimed identity model
Normalised claimed ID model
Cohort model 1
Normalised cohort model
classifier
Cohort model m

Fig. 20