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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
30 June 2016 (30.06.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/105216 A1

(51) International Patent Classification:
G10L 15/20 (2006.01) *G10L 17/20* (2013.01)
G10L 17/02 (2013.01)

(21) International Application Number:
PCT/PL2014/000151

(22) International Filing Date:
22 December 2014 (22.12.2014)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, M/S: RNB-4-150, Santa Clara, CA 95054 (US).

(72) Inventor; and
(71) Applicant : **MAREK, Adam** [PL/PL]; ul. Nieborowska 39/14, PL-80-034 Gdansk (PL).

(72) Inventor: **BOCKLET, Tobias**; Truderinger Str 110, 81673 Munich BY (DE).

(74) Agents: **CZARNOCKA, Oliwia** et al.; Polservice Kancelaria Rzeczników Patentowych sp. z o.o., ul. Bluszczańska 73, PL-00-712 Warszawa (PL).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

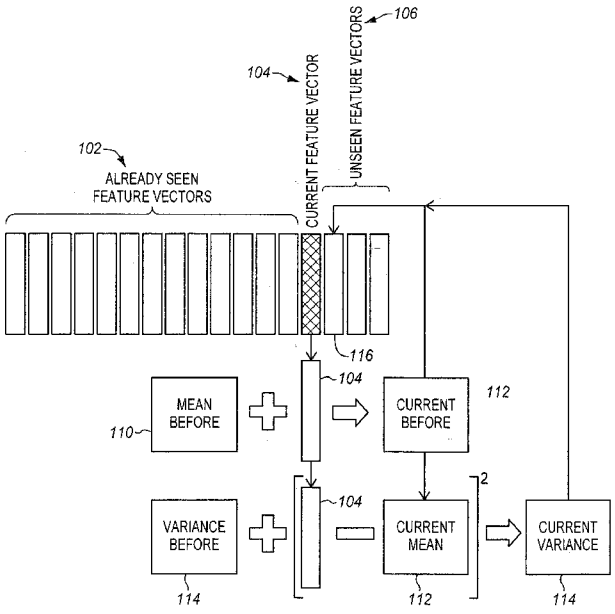
Published:

— *with international search report (Art. 21(3))*

WO 2016/105216 A1

(54) **Title**: CEPSTRAL VARIANCE NORMALIZATION FOR AUDIO FEATURE EXTRACTION

FIG. 1



(57) **Abstract**: Cepstral variance normalization is described for audio feature extraction. In some embodiments a method includes receiving a sequence of frames of digitized audio from a microphone, determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, updating the initial mean to a current mean using the determined feature vector for the first frame, updating the variance to a current variance using the current mean and the determined feature vector for the first frame, determining a next feature vector for each of subsequent frames of the sequence of frames, after determining a next feature vector for each subsequent frame, updating the current mean to a next current mean and updating the current variance to a next current variance and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and sending the determined feature vectors to a recognition stage of an audio processing pipeline.

CEPSTRAL VARIANCE NORMALIZATION FOR AUDIO FEATURE EXTRACTION

FIELD

The present description relates to audio processing and in particular to determining statistical values for cepstral coefficient feature vectors.

BACKGROUND

Voice or speech recognition is increasingly being used as a part of the user interface in computing devices of many different types. Many cellular telephones allow users to push a button and speak into the microphone to perform queries and execute a variety of different commands. Portable and desktop computers perform similar functions. These systems may also convert the speech to text and use this as text input for appointments, messages, or stored documents. Some of these systems process the user speech locally in the device, but many send a recording of the speech to a remote server. Automobiles may also receive voice commands and queries to operate a navigation system or other functions in the vehicle including making calls and sending messages. Voice or speech recognition is also used for identification, logon, and other purposes.

Users prefer a quick response to voice inputs; however, it may require significant processing resources to complete an accurate analysis of the speech or any other audio. In addition many speech analysis techniques are designed to receive a complete utterance and then analyze the complete utterance as a whole. This requires that the system wait for the user to stop speaking and then begin processing using the entire utterance. This inherent latency may be annoying especially when compared to the immediacy of keyboard and mouse inputs. For some systems, there may not be enough memory available to buffer the utterance so that speech analysis is not possible.

One technique used for speech recognition is to analyze the speech for Mel-Frequency Cepstral Coefficients (MFCCs). The MFCC's are compared to a reference for automatic speech recognition (ASR) and speaker recognition. In order to improve the accuracy and reliability of the MFCC, Cepstral Mean Subtraction (CMS) is used in combination with Cepstral Variance Normalization (CVN). CVN can be regarded as noise normalization in that the mean of the signal is subtracted which leads to the removal of stationary noise. These techniques allow the system to be used to good effect not only for user input purposes but also for user authentication and login passwords.

BRIEF DESCRIPTION OF THE DRAWINGS

Embodiments are illustrated by way of example, and not by way of limitation, in the figures of the accompanying drawings in which like reference numerals refer to similar elements.

Figure 1 is a diagram of feature vector calculation for a stream of audio samples in an audio pipeline according to an embodiment.

Figure 2 is a process flow diagram of feature vector determination for a steam of audio samples according to an embodiment.

Figure 3 is a block diagram of a speech recognition system according to an embodiment.

Figure 4 is a block diagram of a computing device incorporating audio processing according to an embodiment.

DETAILED DESCRIPTION

A technique for cepstral variance normalization is described that may be performed in real time locally in a small, portable, or fixed device for verifying the identity of the speaker, understanding a command, and for other applications.

To perform CMS (Cepstral Mean Subtraction) and CVN (Cepstral Variance Normalization) in MFCCs (Mel-Frequency Cepstral Coefficients) the mean and variance of the MFCCs is estimated on a complete audio file. The estimates are then used to apply the normalizations frame-wise on the audio file. As described herein, CVN may be performed as the audio is received sample-by-sample. This may be done without any buffering. Using the described technique, the accuracy of recognition is greatly improved over real-time feature extraction without normalization and is similar to that of offline feature extraction performed in forensics analysis.

A typical technique for determining CMS and CVN offline using the entire audio file may be stated in the form of the equation below.

$$\hat{f}(x_i) = \frac{f(x_i) - \mu}{\sigma} \qquad \text{Eq. 1}$$

where x_i denotes the speech signal at time step i, $f(x_i)$ denotes the feature extraction function, μ denotes the estimated mean of all features, σ denotes the estimated variance. μ and σ are required prior to normalization. This, however, requires a preprocessing of all features of an audio file. As shown the features are normalized by taking the feature for each frame, subtracting the mean and diving by the variance. These are summed for an overall value for the utterance.

The determination of the mean μ over all frames may be done as follows:

$$\mu = \frac{1}{N}\sum_{i=0}^{N-1} f(x_i) \quad \text{Eq. 2}$$

wherein N denotes the number of frames.

Similarly having determined μ, the variance σ of the mean may be determined as follows:

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N-1}(f(x_i) - \mu)^2 \quad \text{Eq. 3}$$

N denotes the number of frames.

In order to determine useful values of the features without first determining a mean and a variance, Equation 1 may be changed as follows:

$$\hat{f}(x_i) = \frac{f(x_i) - \mu_{i-1}}{\sigma_{i-1}} \quad \text{Eq. 4}$$

In this case, the mean and variance values are replaced with an interim mean and variance that is determined up to and including the most recent frame. These values μ_{i-1} and σ_{i-1} are updated for each new frame x_i. The mean and variance after each sample are determined as follows:

$$\mu_i = (1 - \alpha)\mu_{i-1} + \alpha f(x_i) \quad \text{Eq. 5}$$

$$\sigma_i^2 = (1 - \alpha)\sigma_{i-1}^2 + \alpha (f(x_i) - \mu_i)^2 \quad \text{Eq. 6}$$

As shown, each new mean value μ_i is determined by using the previous value μ_{i-1} factored by a parameter α and adjusted by a factored current value of the feature $f(x_i)$. The same approach may be used with the variance, where the current variance σ_i^2 is determined by using the previous value factored by a parameter α and adjusted by a factored current value of the feature $f(x_i)$.

The higher the relevance parameter α is, the more emphasis is put on the actual frame currently being processed, i.e. the i^{th} frame. This technique for in line real time normalization does not introduce any latency or delay as the estimated mean μ and the estimated variance σ are updated with each subsequent frame. The estimated mean and variance rely only on frames that the system has already seen as it processes the audio as the audio is received. There is no look ahead required. This is depicted in Figure 1. In addition, the memory requirements are very small because only the two values for the last frame are used and these are updated with each frame.

Figure 1 is a diagram of feature vector calculation to indicate how feature vectors may be determined as audio is being received. Feature vectors are arriving from right to left with the most recently received feature vectors on the right. The current mean 112 and the current

variance 114 are used to determine the incoming feature vector 116. As the feature vectors move to the left, the incoming feature vector 116 becomes the current feature vector 104. The current feature vector 104 is then used to update the prior mean 110 to determine the current mean 112. Similarly, the prior variance 114 is used together with the new current mean 112 to determine the current variance 114. The current mean and variance are then plugged into determining or processing the upcoming unused feature vector 116.

The already seen feature vectors 102 increase as the current feature vector 104 moves to the left and unseen feature vectors 106 become the current feature vector 104 and then the already seen feature vectors 102. The feature vectors may be summed or analyzed as they are produced so that the analysis may be finished almost immediately after the end of the utterance.

The same principles set forth in Equations 4, 5, and 6 may also be expressed in pseudo code. In the example below "feat" denotes a feature vector.

```
if((numberFrame == 0) && (init) ){

   for(j=0;j<feat.length;j++){
      mean[j] = initMean[j];
      sigsq[j] =  initsig[j];
   }


} else if( (numberFrame == 0) && (!init)){

   for(j=0;j<feat.length;j++){
      mean[j] = feat[j];
      sigsq[j] =  feat[j] * feat[j];
   }

} else {

   for(j=0;j<feat.length;j++){
      mean[j] = (mean[j]*(1-alpha)) + (alpha * feat[j]);
      sigsq[j] = sigsq[j]*(1-alpha) +
                  (alpha * ((feat[j]-mean[j]) * (feat[j]-
mean[j])) );
   }

}

   for (j = 0; j < feat.length; j++) {
               feat[j] -= mean[j];
               feat[j] /= Math.sqrt(sigsq[j]);
   }

numberFrame++;
```

As shown, the whole audio file need not be read and buffered. The memory requirements latency from buffering the whole audio stream is avoided. This is particularly valuable for using speaker recognition for logins and for voice authentication.

The described CVN may be used to improve an online feature extraction system. The described online CMS and CVN may be performed without introducing any latency. The CVN does not require any offline processing and estimates the variance based on already seen data. During runtime, the variance is updated with every new feature frame.

The online CVN provides a performance in voice recognition that is similar to conventional offline techniques without requiring any form of feature buffering and without producing latency because the system does not require any knowledge of future feature frames.

The online techniques herein may be used with only CMS or CMS may be combined with CVN. A further modification may be made by initializing the mean μ and the variance σ to some value before the processing is started. Initial values for mean and variance can be specified by estimating appropriate values empirically using similar hardware. This may provide a significant improvement for determining early features. However, the advantages of the initial values are diminished as more features vectors are received.

Figure 2 is a process flow diagram of determining a CMS and a CVN online or in real-time as audio frames arrive. The system starts based on receiving audio frames from a microphone or other source. The audio may be processed or formatted in any of a variety of different ways. Typically the audio is in a digital sample format. At 202, the CMS (μ) value is optionally initialized and at 204, the CVN (σ) values are optionally initialized. These may be initialized to zero, to a middle value, or to some accurate estimate of a typical value for the particular type of audio and the hardware system that is receiving the audio.

At 206, a feature vector is determined for the first frame in the received audio. This may be done as described in the context of Equation 4. The initialization of the mean and variance of 202 is optional. If there is no initialization then the values from the first frame's feature vector may be used. In other words the mean μ is set equal to the feature value and the variance will be zero because there is only one mean value. For the next frame the same values from the first frame's feature vector may be used again as the values for the previous frame. As a further alternative, if there is no initialization then 0 may be used for μ and 1 may be used for σ. The second frame will then use the values from the first frame. The determination of the first frame's feature vector may be performed before the last frame has been received. The feature vector processing may start as soon as the first frame is received.

At 208 μ is updated using the determined feature vector. If there is no initialization and this is the first frame then updating μ corresponds to establishing μ in the first instance using the

feature value for the first frame. At 210 σ is updated using the updated μ and the determined feature vector. If μ and σ were initialized then the initial values may be included in this updating process. An example updating process is set forth in Equations 5 and 6. If there is no initialized value, then the values of the first frame may be used as initial values, e.g. μ=(feature vector) and σ=(feature vector)*(feature vector). So equations 5 and 6 will be processed with these initial values for the first frame. If instead there are no initial values at all, that is if there is no current value of μ and σ to be updated, then the first terms in Equations 5 and 6 drop out for the first frame.

At 212, the feature vector for the next frame is determined using the updated μ and σ values. This process of updating μ and σ and determining feature vectors continues until the entire utterance is processed. At 214, it is determined whether there is a next frame. The operations continue until all of the frames are processed. At that time the feature vectors are sent to the recognition stage (216) and the process ends.

Figure 3 is a block diagram of a speech recognition system 300 which may use feature vectors and CMS and CVN values determined as described herein. While embodiments are described in the context of a speech recognition system, the approach described herein and in the context of Figure 3 is not restricted to speech recognition, but may also be used with speaker recognition or authentication, language recognition, speaker diarization, etc. An input signal, such as speech from a microphone, is received at an analog to digital converter (A/D) 310. A feature extractor or spectral analysis unit 320 receives digitized speech frames from the A/D. A decoder 330 compares the extracted features received from the feature extractor to a connected acoustic model 340, and a connected language model 350 and produces recognized speech 360 or commands as an output.

The input signal 305 represents the input speech or some other audio. It is first digitized using the A/D 310. The digital signal is then sliced up into frames typically of 10, 15, or 20 ms in the A/D or the feature extractor. Each frame of the signal is then converted into a corresponding feature vector which is used to characterize the spectral properties of the input signal.

The feature vector may have one or more dimensions and may include one or more different components. These components may include, for example MFCC components, delta MFCCs, and delta-delta MFCCs. The feature vectors generated by the feature extractor unit 320 are then submitted into the decoder 330 which determines whether the speech matches an authenticated user with enough accuracy. Alternatively, it may determine a sentence or a sequence of words that has the highest probability given the acoustic events characterized by the feature vectors. This is done using any of a variety of different models 340, 350 or analysis.

The recognized speech 360 may then be provided for storage, execution, or communication to others.

Figure 4 illustrates a computing device 100 in accordance with one implementation. The computing device 100 houses a system board 2. The board 2 may include a number of components, including but not limited to a processor 4 and at least one communication package 6. The communication package is coupled to one or more antennas 16. The processor 4 is physically and electrically coupled to the board 2.

Depending on its applications, computing device 100 may include other components that may or may not be physically and electrically coupled to the board 2. These other components include, but are not limited to, volatile memory (e.g., DRAM) 8, non-volatile memory (e.g., ROM) 9, flash memory (not shown), a graphics processor 12, a digital signal processor (not shown), a crypto processor (not shown), a chipset 14, an antenna 16, a display 18 such as a touchscreen display, a touchscreen controller 20, a battery 22, an audio codec (not shown), a video codec (not shown), a power amplifier 24, a global positioning system (GPS) device 26, a compass 28, an accelerometer (not shown), a gyroscope (not shown), a speaker 30, a camera 32, a microphone array 34, and a mass storage device (such as hard disk drive) 10, compact disk (CD) (not shown), digital versatile disk (DVD) (not shown), and so forth). These components may be connected to the system board 2, mounted to the system board, or combined with any of the other components.

The communication package 6 enables wireless and/or wired communications for the transfer of data to and from the computing device 100. The term "wireless" and its derivatives may be used to describe circuits, devices, systems, methods, techniques, communications channels, etc., that may communicate data through the use of modulated electromagnetic radiation through a non-solid medium. The term does not imply that the associated devices do not contain any wires, although in some embodiments they might not. The communication package 6 may implement any of a number of wireless or wired standards or protocols, including but not limited to Wi-Fi (IEEE 802.11 family), WiMAX (IEEE 802.16 family), IEEE 802.20, long term evolution (LTE), Ev-DO, HSPA+, HSDPA+, HSUPA+, EDGE, GSM, GPRS, CDMA, TDMA, DECT, Bluetooth, Ethernet derivatives thereof, as well as any other wireless and wired protocols that are designated as 3G, 4G, 5G, and beyond. The computing device 100 may include a plurality of communication packages 6. For instance, a first communication package 6 may be dedicated to shorter range wireless communications such as Wi-Fi and Bluetooth and a second communication package 6 may be dedicated to longer range wireless communications such as GPS, EDGE, GPRS, CDMA, WiMAX, LTE, Ev-DO, and others.

The microphones 34 and the speaker 30 are coupled to one or more audio chips 36 to perform digital conversion, feature extraction, encoding and decoding, and frame based processing as described herein. The processor 4 is coupled to the audio chip, through an audio front end, for example, to drive the process with interrupts, set parameters, and control operations of the audio chip. Frame-based processing may be performed in the audio chip or in the communication package 6. Power management functions may be performed by the processor, coupled to the battery 22 or a separate power management chip may be used.

In various implementations, the computing device 100 may be a laptop, a netbook, a notebook, an ultrabook, a smartphone, a wearable device, a tablet, a personal digital assistant (PDA), an ultra mobile PC, a mobile phone, a desktop computer, a server, a printer, a scanner, a monitor, a set-top box, an entertainment control unit, a digital camera, a portable music player, or a digital video recorder. The computing device may be fixed, portable, or wearable. In further implementations, the computing device 100 may be any other electronic device that processes data.

Embodiments may be implemented as a part of one or more memory chips, controllers, CPUs (Central Processing Unit), microchips or integrated circuits interconnected using a motherboard, an application specific integrated circuit (ASIC), and/or a field programmable gate array (FPGA).

References to "one embodiment", "an embodiment", "example embodiment", "various embodiments", etc., indicate that the embodiment(s) so described may include particular features, structures, or characteristics, but not every embodiment necessarily includes the particular features, structures, or characteristics. Further, some embodiments may have some, all, or none of the features described for other embodiments.

In the following description and claims, the term "coupled" along with its derivatives, may be used. "Coupled" is used to indicate that two or more elements co-operate or interact with each other, but they may or may not have intervening physical or electrical components between them.

As used in the claims, unless otherwise specified, the use of the ordinal adjectives "first", "second", "third", etc., to describe a common element, merely indicate that different instances of like elements are being referred to, and are not intended to imply that the elements so described must be in a given sequence, either temporally, spatially, in ranking, or in any other manner.

The drawings and the forgoing description give examples of embodiments. Those skilled in the art will appreciate that one or more of the described elements may well be combined into a single functional element. Alternatively, certain elements may be split into multiple functional elements. Elements from one embodiment may be added to another embodiment. For example,

orders of processes described herein may be changed and are not limited to the manner described herein. Moreover, the actions of any flow diagram need not be implemented in the order shown; nor do all of the acts necessarily need to be performed. Also, those acts that are not dependent on other acts may be performed in parallel with the other acts. The scope of embodiments is by no means limited by these specific examples. Numerous variations, whether explicitly given in the specification or not, such as differences in structure, dimension, and use of material, are possible. The scope of embodiments is at least as broad as given by the following claims.

The following examples pertain to further embodiments. The various features of the different embodiments may be variously combined with some features included and others excluded to suit a variety of different applications. Some embodiments pertain to a method that includes receiving a sequence of frames of digitized audio from a microphone, determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, updating the initial mean to a current mean using the determined feature vector for the first frame, updating the variance to a current variance using the current mean and the determined feature vector for the first frame, determining a next feature vector for each of subsequent frames of the sequence of frames, after determining a next feature vector for each subsequent frame, updating the current mean to a next current mean and updating the current variance to a next current variance and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and sending the determined feature vectors to a recognition stage of an audio processing pipeline.

Further embodiments include setting the initial mean and variance to estimated values for the digitized audio.

In further embodiments the initial mean and variance are set based on the feature vector determined for the first frame.

In further embodiments determining a feature vector comprises determining a mel-frequency cepstral coefficient.

In further embodiments the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

In further embodiments determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

In further embodiments the digitized audio is speech.

In further embodiments the recognition stage comprises a decoder, the method further comprising receiving the determined feature vectors at the decoder and determining at the decoder whether the digitized speech matches an authenticated user.

In further embodiments the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

In further embodiments the next current variance is determined after each next feature vector using only the corresponding next current mean.

Some embodiments pertain to a machine-readable medium having instructions thereon that, when executed by the machine, cause the machine to perform operations that include receiving a sequence of frames of digitized audio from a microphone, determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, updating the initial mean to a current mean using the determined feature vector for the first frame, updating the variance to a current variance using the current mean and the determined feature vector for the first frame, determining a next feature vector for each of subsequent frames of the sequence of frames, after determining a next feature vector for each subsequent frame, updating the current mean to a next current mean and updating the current variance to a next current variance and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and sending the determined feature vectors to a recognition stage of an audio processing pipeline.

In further embodiments determining a feature vector comprises determining a mel-frequency cepstral coefficient.

In further embodiments the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

In further embodiments determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

In further embodiments the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

In further embodiments the next current variance is determined after each next feature vector using only the corresponding next current mean.

Some embodiments pertain to an apparatus that includes a feature extractor to receive a sequence of frames of digitized audio from a microphone through an analog to digital converter, to determine a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, to update the initial mean to a current mean using the determined feature vector for the first frame, to update the variance to a current variance using the current mean and the determined feature vector for the first frame, to determine a next feature vector for each of subsequent frames of the sequence of frames, to

update the current mean to a next current mean and update the current variance to a next current variance after determining a next feature vector for each subsequent frame, wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and a recognition stage to recognize the extracted feature vectors.

In further embodiments the digitized audio is speech.

In further embodiments the recognition stage comprises a decoder, the decoder receiving the determined feature vectors and determining whether the digitized speech matches an authenticated user.

In further embodiments the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

Some embodiments pertain to an apparatus that includes means for receiving a sequence of frames of digitized audio from a microphone, means for determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, means for updating the initial mean to a current mean using the determined feature vector for the first frame, means for updating the variance to a current variance using the current mean and the determined feature vector for the first frame, means for determining a next feature vector for each of subsequent frames of the sequence of frames, for updating the current mean to a next current mean and updating the current variance to a next current variance after determining a next feature vector for each subsequent frame and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and means for sending the determined feature vectors to a recognition stage of an audio processing pipeline.

In further embodiments the initial mean and variance are set based on the feature vector determined for the first frame.

In further embodiments determining a feature vector comprises determining a mel-frequency cepstral coefficient.

In further embodiments the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

In further embodiments determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

In further embodiments the recognition stage comprises means for decoding, the means for decoding determining whether the digitized speech matches an authenticated user.

Some embodiments pertain to a computing system that includes a plurality of microphones to receive speech, an analog to digital converter to convert the speech to a sequence

of frames of digitized audio, a feature extractor to receive the sequence of frames of digitized audio from the analog to digital converter, to determine a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, to update the initial mean to a current mean using the determined feature vector for the first frame, to update the variance to a current variance using the current mean and the determined feature vector for the first frame, to determine a next feature vector for each of subsequent frames of the sequence of frames, to update the current mean to a next current mean and update the current variance to a next current variance after determining a next feature vector for each subsequent frame, wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance, and a recognition stage to recognize the extracted feature vectors.

In further embodiments decoder, the decoder receiving the determined feature vectors and determining whether the digitized speech matches an authenticated user.

CLAIMS

1. A method comprising:

receiving a sequence of frames of digitized audio from a microphone;

determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance;

updating the initial mean to a current mean using the determined feature vector for the first frame;

updating the variance to a current variance using the current mean and the determined feature vector for the first frame;

determining a next feature vector for each of subsequent frames of the sequence of frames;

after determining a next feature vector for each subsequent frame, updating the current mean to a next current mean and updating the current variance to a next current variance and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance; and

sending the determined feature vectors to a recognition stage of an audio processing pipeline.

2. The method of Claim 1, further comprising setting the initial mean and variance to estimated values for the digitized audio.

3. The method of Claim 1 or 2, wherein the initial mean and variance are set based on the feature vector determined for the first frame.

4. The method of any one or more of the above claims, wherein determining a feature vector comprises determining a mel-frequency cepstral coefficient.

5. The method of Claim 4, wherein the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

6. The method of Claim 5, wherein determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

7. The method of any one or more of the above claims, wherein the digitized audio is speech.

8. The method of any one or more of the above claims, wherein the recognition stage comprises a decoder, the method further comprising receiving the determined feature vectors at the decoder and determining at the decoder whether the digitized speech matches an authenticated user.

9. The method of any one or more of the above claims, wherein the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

10. The method of Claim 9, wherein the next current variance is determined after each next feature vector using only the corresponding next current mean.

11. A machine-readable medium having instructions thereon that, when executed by the machine, cause the machine to perform operations comprising:

receiving a sequence of frames of digitized audio from a microphone;

determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance;

updating the initial mean to a current mean using the determined feature vector for the first frame;

updating the variance to a current variance using the current mean and the determined feature vector for the first frame;

determining a next feature vector for each of subsequent frames of the sequence of frames;

after determining a next feature vector for each subsequent frame, updating the current mean to a next current mean and updating the current variance to a next current variance and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance; and

sending the determined feature vectors to a recognition stage of an audio processing pipeline.

12. The medium of Claim 11, wherein determining a feature vector comprises determining a mel-frequency cepstral coefficient.

13. The medium of Claim 12, wherein the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

14. The medium of Claim13, wherein determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

15. The medium of any one or more of claims 11-14, wherein the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

16. The medium of Claim 15, wherein the next current variance is determined after each next feature vector using only the corresponding next current mean.

17. An apparatus comprising:

a feature extractor to receive a sequence of frames of digitized audio from a microphone through an analog to digital converter, to determine a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance, to update the initial mean to a current mean using the determined feature vector for the first frame, to update the variance to a current variance using the current mean and the determined feature vector for the first frame, to determine a next feature vector for each of subsequent frames of the sequence of frames, to update the current mean to a next current mean and update the current variance to a next current variance after determining a next feature vector for each subsequent frame, wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance; and

a recognition stage to recognize the extracted feature vectors.

18. The apparatus of Claim 17, wherein the digitized audio is speech.

19. The apparatus of Claim 17 or 18, wherein the recognition stage comprises a decoder, the decoder receiving the determined feature vectors and determining whether the digitized speech matches an authenticated user.

20. The apparatus of Claim 17, 18, or 19, wherein the next current mean is determined after each next feature vector using only the corresponding next feature vector and feature vectors of earlier frames of the sequence of frames.

21. An apparatus comprising:

means for receiving a sequence of frames of digitized audio from a microphone;

means for determining a feature vector for a first frame of the sequence of frames, the feature vector being determined using an initial mean and an initial variance;

means for updating the initial mean to a current mean using the determined feature vector for the first frame;

means for updating the variance to a current variance using the current mean and the determined feature vector for the first frame;

means for determining a next feature vector for each of subsequent frames of the sequence of frames, for updating the current mean to a next current mean and updating the current variance to a next current variance after determining a next feature vector for each subsequent frame and wherein determining a feature vector for a subsequent frame comprises using the next current mean and the next current variance; and

means for sending the determined feature vectors to a recognition stage of an audio processing pipeline.

22. The apparatus of Claim 21, wherein the initial mean and variance are set based on the feature vector determined for the first frame.

23. The apparatus of Claim 21 or 22, wherein determining a feature vector comprises determining a mel-frequency cepstral coefficient.

24. The apparatus of Claim 23, wherein the current mean comprises a cepstral mean and the current variance comprises a cepstral variance.

25. The apparatus of Claim 24, wherein determining a next feature vector comprises determining a next feature vector using cepstral mean substraction and cepstral variance normalization.

26. The apparatus of any one or more of claims 21-25, wherein the recognition stage comprises means for decoding, the means for decoding determining whether the digitized speech matches an authenticated user.

FIG. 1

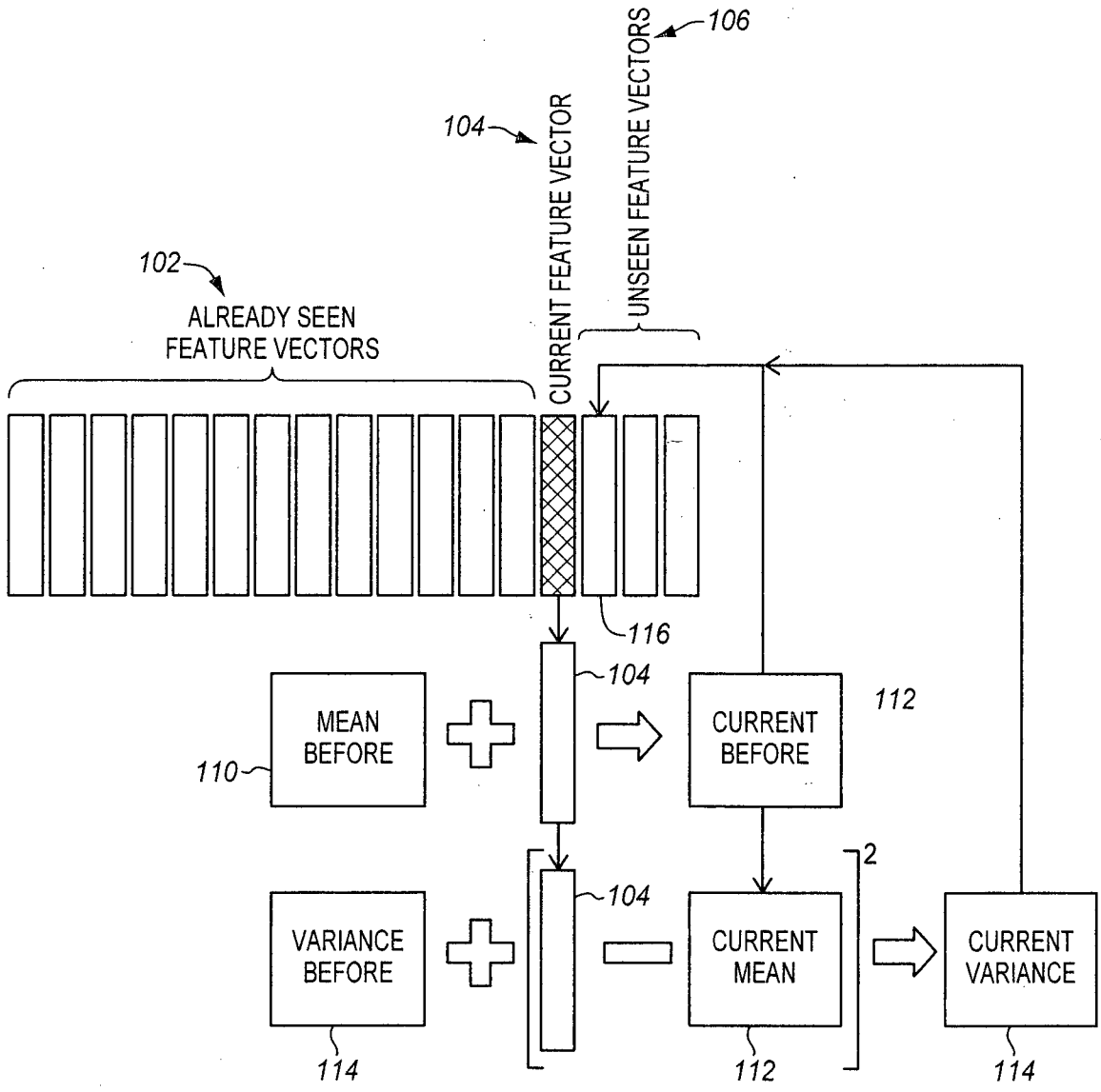
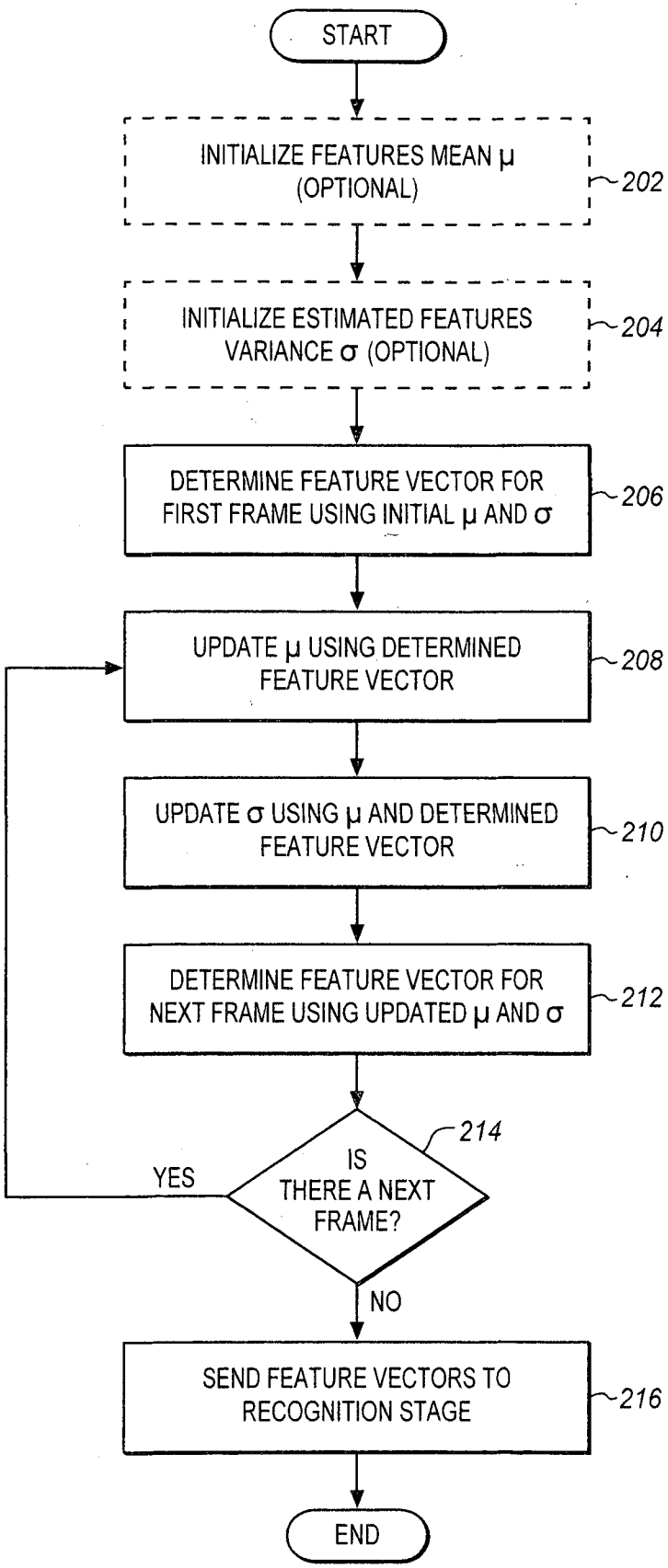

106
UNSEEN FEATURE VECTORS
104
CURRENT FEATURE VECTOR
102
ALREADY SEEN FEATURE VECTORS
116
104
MEAN BEFORE
110
CURRENT BEFORE
112
2
104
VARIANCE BEFORE
CURRENT MEAN
CURRENT VARIANCE
114
112
114

FIG. 2


START
INITIALIZE FEATURES MEAN μ (OPTIONAL)
202
INITIALIZE ESTIMATED FEATURES VARIANCE σ (OPTIONAL)
204
DETERMINE FEATURE VECTOR FOR FIRST FRAME USING INITIAL μ AND σ
206
UPDATE μ USING DETERMINED FEATURE VECTOR
208
UPDATE σ USING μ AND DETERMINED FEATURE VECTOR
210
DETERMINE FEATURE VECTOR FOR NEXT FRAME USING UPDATED μ AND σ
212
IS THERE A NEXT FRAME?
214
YES
NO
SEND FEATURE VECTORS TO RECOGNITION STAGE
216
END

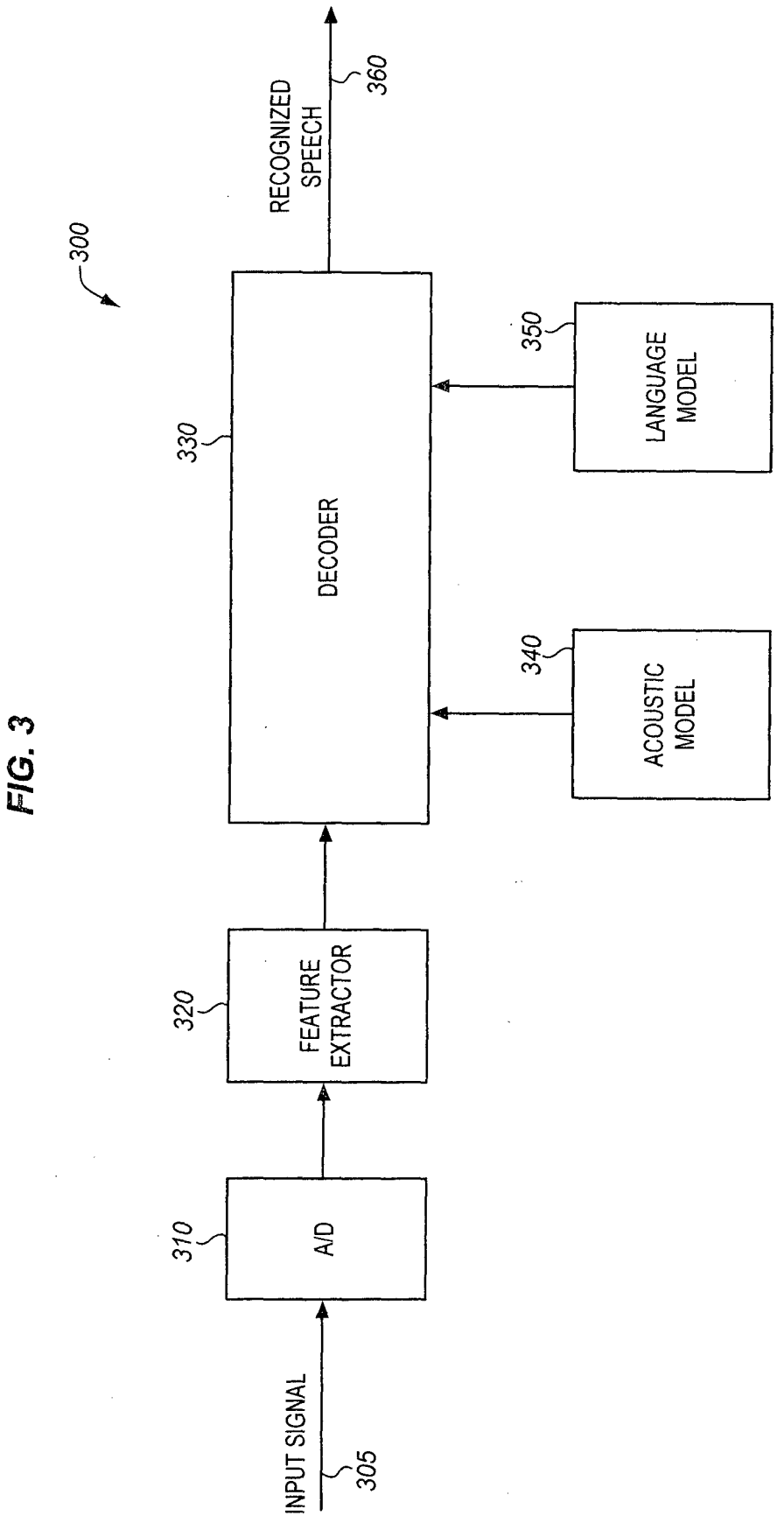
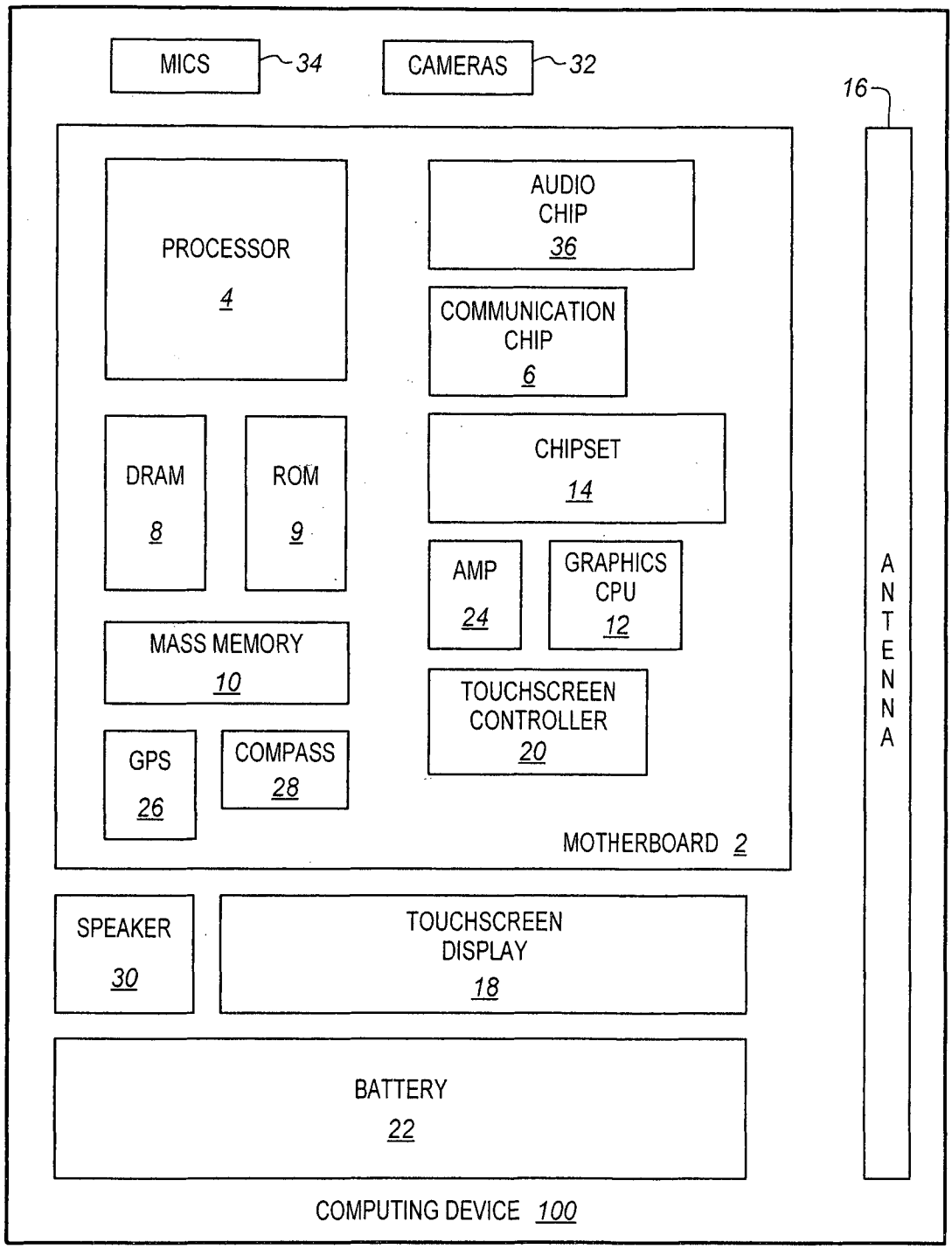

300
INPUT SIGNAL
305
A/D
310
FEATURE EXTRACTOR
320
DECODER
330
ACOUSTIC MODEL
340
LANGUAGE MODEL
350
RECOGNIZED SPEECH
360

FIG. 3

FIG. 4


MICS
34
CAMERAS
32
16
PROCESSOR
4
AUDIO CHIP
36
COMMUNICATION CHIP
6
DRAM
8
ROM
9
CHIPSET
14
AMP
24
GRAPHICS CPU
12
MASS MEMORY
10
TOUCHSCREEN CONTROLLER
20
GPS
26
COMPASS
28
MOTHERBOARD 2
ANTENNA
SPEAKER
30
TOUCHSCREEN DISPLAY
18
BATTERY
22
COMPUTING DEVICE 100

INTERNATIONAL SEARCH REPORT

International application No
PCT/PL2014/000151

A. CLASSIFICATION OF SUBJECT MATTER
INV. G10L15/20 G10L17/02 G10L17/20
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, INSPEC, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PUJOL P ET AL: "On Real-Time Mean-and-Variance Normalization of Speech Recognition Features", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 2006. ICASSP 2006 PROCEEDINGS . 2006 IEEE INTERNATIONAL CONFERENCE ON TOULOUSE, FRANCE 14-19 MAY 2006, PISCATAWAY, NJ, USA,IEEE, PISCATAWAY, NJ, USA, 1 January 2006 (2006-01-01), pages I-I, XP031100404, ISBN: 978-1-4244-0469-8 page 773, left-hand column, line 1 - page 776, right-hand column, line 25; tables 3,4 ----- -/--	1-26

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 June 2015	19/06/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Dobler, Ervin

INTERNATIONAL SEARCH REPORT

International application No
PCT/PL2014/000151

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VIIKKI O ET AL: "A recursive feature vector normalization approach for robust speech recognition in noise", ACOUSTICS, SPEECH AND SIGNAL PROCESSING, 1998. PROCEEDINGS OF THE 1998 IEEE INTERNATIONAL CONFERENCE ON SEATTLE, WA, USA 12-15 MAY 1998, NEW YORK, NY, USA,IEEE, US, vol. 2, 12 May 1998 (1998-05-12), pages 733-736, XP010279266, DOI: 10.1109/ICASSP.1998.675369 ISBN: 978-0-7803-4428-0 page 734, left-hand column, line 3 - page 735, left-hand column, last line -----	1-26
X	US 2005/182621 A1 (ZLOKARNIK IGOR [US] ET AL) 18 August 2005 (2005-08-18) paragraph [0002] - paragraph [0007] -----	1-7, 9-18, 20-25

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/PL2014/000151

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005182621 A1	18-08-2005	CN 101228577 A	23-07-2008
		EP 1774516 A2	18-04-2007
		JP 4682154 B2	11-05-2011
		JP 2007536562 A	13-12-2007
		US 2005182621 A1	18-08-2005
		WO 2005070130 A2	04-08-2005