



(51) International Patent Classification:

G10L 15/16 (2006.01)

(21) International Application Number:

PCT/US2017/043765

(22) International Filing Date:

25 July 2017 (25.07.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(71) Applicant: HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P. [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).

(72) Inventors: BHARITKAR, Sunil; 1501 Page Mill Road, Palo Alto, California 94304 (US). MURPHY, David; 1501 Page Mill Road, Palo Alto, California 94304 (US).

(74) Agent: BURROWS, Sarah E. et al.; HP Inc., 3390 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 80528 (US).

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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).

(54) Title: LANGUAGE IDENTIFICATION WITH SPEECH AND VISUAL ANTHROPOMETRIC FEATURES

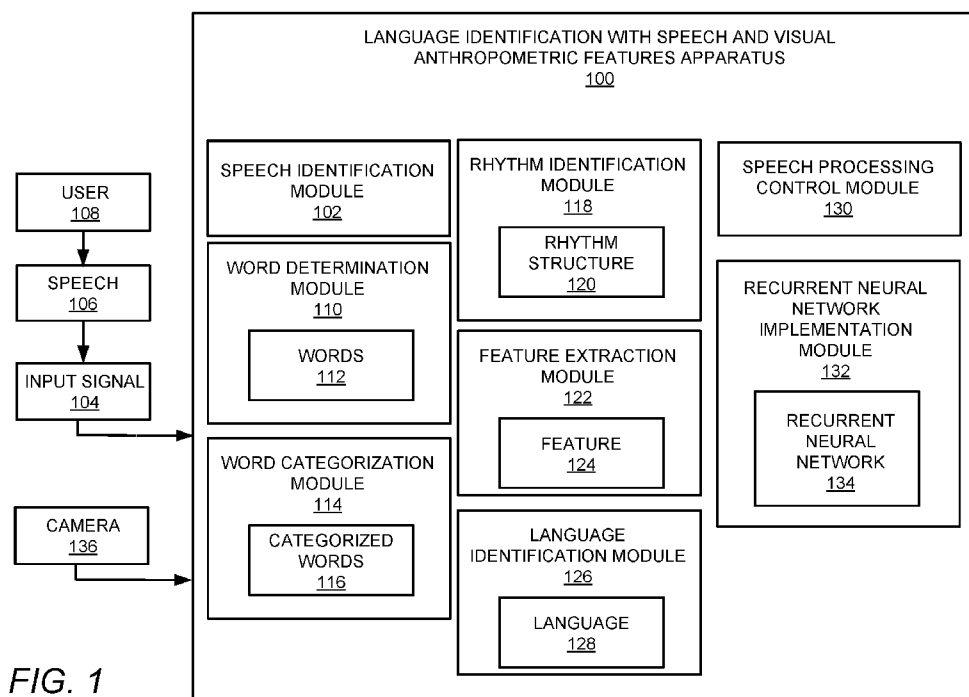


FIG. 1

(57) Abstract: In some examples, language identification with speech and visual anthropometric features may include analyzing an input signal to identify a speech of a user, and extracting a feature of the speech. Language identification with speech and visual anthropometric features may further include identifying, based on a classification of the feature, a language of the speech, and controlling processing associated with the speech based on the identified language.

Declarations under Rule 4.17:

- *as to the identity of the inventor (Rule 4.17(i))*
- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

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

LANGUAGE IDENTIFICATION WITH SPEECH AND VISUAL ANTHROPOMETRIC FEATURES

BACKGROUND

[0001] Speech may be translated into text or another form. For example, automatic speech recognition (ASR) may be used to transcribe speech into readable text in real time. Some speech recognition systems may need for a user to read text or isolated vocabulary into the system. The speech recognition system may analyze the user's specific voice, and use attributes of the user's specific voice to fine-tune the recognition of that user's speech, resulting in increased accuracy.

BRIEF DESCRIPTION OF DRAWINGS

[0002] Features of the present disclosure are illustrated by way of example and not limited in the following figure(s), in which like numerals indicate like elements, in which:

[0003] Figure 1 illustrates an example layout of a language identification with speech and visual anthropometric features apparatus;

[0004] Figure 2 illustrates an example of language identification using speech to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0005] Figure 3 illustrates an example of language identification using speech and camera input to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0006] Figure 4 illustrates an example of feature based training and transfer of a neural network model to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0007] Figure 5 illustrates an example of a multilayer feedforward neural network to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0008] Figure 6 illustrates English emulated "common" spoken word (family) by a female, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0009] Figure 7 illustrates Russian emulated "common" spoken word (sem'ya for family) by a male, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the language identification with speech and

visual anthropometric features apparatus of Figure 1;

[0010] Figure 8 illustrates Russian emulated "common" spoken word (sem'ya for family) by a female, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0011] Figure 9 illustrates processed female English tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 6, to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0012] Figure 10 illustrates processed male Russian tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 7, to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0013] Figure 11 illustrates processed female Russian tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 8, to illustrate operation of the language identification with speech and visual anthropometric features apparatus of Figure 1;

[0014] Figure 12 illustrates an example block diagram for language identification with speech and visual anthropometric features;

[0015] Figure 13 illustrates an example flowchart of a method for language identification with speech and visual anthropometric features; and

[0016] Figure 14 illustrates a further example block diagram for language identification with speech and visual anthropometric features.

DETAILED DESCRIPTION

[0017] For simplicity and illustrative purposes, the present disclosure is described by referring mainly to examples. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. It will be readily apparent however, that the present disclosure may be practiced without limitation to these specific details. In other instances, some methods and structures have not been described in detail so as not to unnecessarily obscure the present disclosure.

[0018] Throughout the present disclosure, the terms "a" and "an" are intended to denote at least one of a particular element. As used herein, the term "includes" means includes but not limited to, the term "including" means including but not limited to. The term "based on" means based at least in part on.

[0019] Apparatuses for language identification with speech and visual anthropometric features, methods for language identification with speech and visual anthropometric features, and non-transitory computer readable media having stored thereon machine readable instructions to provide language identification with speech and visual anthropometric features are disclosed. The apparatuses, methods, and non-transitory computer readable media disclosed herein provide for automatic language identification during speech-based communication between people. For example, in a collaboration setting (e.g., using a smart-surface) or a conference call, languages that are being spoken may be identified at the start of a conversation so as to enable linguistically-relevant speech-synthesis via a machine-translation schema. In addition to speech, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide for detection of demographic-based visual features of a speaker, for example, by a camera, to create a multimodal approach for language identification. Thus, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide for use cases such as speech to text translation, communication between participants speaking different languages, etc.

[0020] With respect to language identification techniques, it is technically challenging to sense and recognize speech to determine what language is being spoken. It is also technically challenging to sense and recognize speech with a high degree of confidence, and/or to combine multiple techniques of sensing and recognizing speech.

[0021] In order to address at least these technical challenges associated with language identification, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide for language identification that includes, for example, word boundary detection in the speech, identification of a rhythm structure of the speech, and/or feature identification, where these aspects are used in combination to identify a language being spoken. Further, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide for language identification that includes feature identification based on fundamental frequency analysis of words.

[0022] According to an example, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide a database based approach for which word boundary identification is performed on speech. Thereafter, a correlation may be performed with respect to the fundamental atoms of the words in the database that is available for various languages and various frequently used words. Based on the results (e.g., what language, and associated probability), the spoken language of the speech may be detected. The features that may be extracted from speech and analyzed may include, for example, the fundamental frequency, spectral centroid, how words evolve in time, etc. The database approach may also employ a neural network, where features may be stored for various languages in the database, and the neural network may be pre-trained on certain ones of the features. Thus, when certain features are present in speech, the neural network may classify the language according to the pre-trained features that are frequently occurring based on key areas (e.g., fundamental frequency, spectral centroid, how words evolve in time, etc.), or words.

[0023] According to an example, the apparatuses, methods, and non-transitory computer readable media disclosed herein provide a recurrent neural network based approach, where the recurrent neural network may be trained on various examples of spoken languages by different speakers. Having trained the recurrent neural network, for new speech, the recurrent neural network may identify which language is being spoken, and redirect the results to the correct automatic speech recognition (ASR) to have the speech interpreted properly. The features in this regard may be extracted from the speech based on temporal signal.

[0024] For the apparatuses, methods, and non-transitory computer readable media disclosed herein, modules, as described herein, may be any combination of hardware and programming to implement the functionalities of the respective modules. In some examples described herein, the combinations of hardware and programming may be implemented in a number of different ways. For example, the programming for the modules may be processor executable instructions stored on a non-transitory machine-readable storage medium and the hardware for the modules may include a processing resource to execute those instructions. In these examples, a computing device implementing such modules may include the machine-readable storage medium storing the instructions and the processing resource to execute the instructions, or the machine-readable storage medium may be separately stored and accessible by the computing device and the processing resource. In some examples, some modules may be implemented in circuitry.

[0025] Figure 1 illustrates an example layout of a language identification with speech and visual anthropometric features apparatus (hereinafter also referred to as “apparatus 100”).

[0026] Referring to Figure 1, the apparatus 100 may include a speech identification module 102 to analyze an input signal 104 to identify a speech 106 of a user 108. For example, the speech identification module 102 is to analyze the input signal 104 to identify the speech 106 of the user 108 by performing, based on a zero-crossing rate and/or a periodicity analysis, a voice activity detection.

[0027] A word determination module 110 is to determine, based on word boundary detection, a plurality of words 112 included in the speech 106.

[0028] A word categorization module 114 is to categorize the words 112, for example, into a phoneme, a vowel, and/or a consonant, to generate categorized words 116.

[0029] A rhythm identification module 118 is to identify a rhythm structure 120 of the speech 106. For example, the rhythm identification module 118 is to identify the rhythm structure 120 of the speech 106 by determining a vowel duration of a vowel included in the words 112, and analyzing, based on the vowel duration, a variability in a length of the vowel included in the words 112.

[0030] A feature extraction module 122 is to extract a feature 124 of the speech 106.

[0031] A language identification module 126 is to identify, based on a classification of the categorized words 112, the identified rhythm structure 120, and the extracted feature 124, a language 128 of the speech 106.

[0032] According to an example, the feature extraction module 122 is to analyze landmark features of the user 108 to determine a racial characteristic of the user 108, and the language identification module 126 is to combine results including the language identification and the determined racial characteristic to increase a confidence associated with the identification of the language 128 of the speech 106. For example, the feature extraction module 122 is to analyze landmark features of the user 108 as received from an image captured by a camera 136.

[0033] According to an example, the feature extraction module 122 is to classify, based on an analysis of an image of the user by a pre-trained learning model, the user according to an ethnicity, and the language identification module 126 is to combine results including the language identification and the classified ethnicity to increase a confidence associated with the identification of the language

128 of the speech 106.

[0034] According to an example, the feature extraction module 122 is to extract the feature 124 of the speech 106 by analyzing a fundamental frequency f_0 of a word of the plurality of words 112 by a trained learning model.

[0035] A speech processing control module 130 is to control processing associated with the speech 106 based on the identified language 128. For example, the speech processing control module 130 is to control processing associated with the speech 106 based on the identified language by implementing, for the identified language 128, a speech to text translation, and/or translating the identified language 128 to another language specific to another user.

[0036] According to an example, instead of or in combination with operation of the word determination module 110, the word categorization module 114, the rhythm identification module 118, and/or the feature extraction module 122, a recurrent neural network implementation module 132 is to analyze, by a trained recurrent neural network 134, the speech 106 of the user 108. The recurrent neural network implementation module 132 is to extract, based on the analysis of the speech 106 by the recurrent neural network 134, features of the speech 106. The language identification module 126 is to identify, based on a classification of the features by the recurrent neural network 134, a language 128 of the speech 106. Further, the speech processing control module 130 is to control processing associated with the speech 106 based on the identified language 128.

[0037] Figure 2 illustrates an example of language identification using speech to illustrate operation of the apparatus 100.

[0038] Referring to Figure 2, at block 200 the speech identification module 102 is to analyze the input signal 104 to identify the speech 106 of the user 108 by performing, based on a zero-crossing rate and/or a periodicity analysis, a voice activity detection. For example, the speech identification module 102 is to analyze the input signal 104 to identify the speech 106 of the user 108 by performing,

based on a zero-crossing rate, periodicity analysis (e.g., fundamental and formant frequencies f_0 , f_1 , f_2), centroid, critical-band filter coefficients, mel-frequency cepstrum (MFC) coefficients, etc., a voice activity detection. The voice activity detection may also include noise suppression. For example, the noise suppression may be performed by implementing spectral subtraction, and may be used in conjunction with the voice activity detection. The noise suppression may also include a blind de-reverberation technique to reduce the influence of strong acoustical reflections that impede proper speech recognition.

[0039] At block 202, the word determination module 110 is to determine, based on word boundary detection, the plurality of words 112 included in the speech 106.

[0040] At block 204, the feature extraction module 122 is to extract a feature 124 of the speech 106.

[0041] At block 206, the language identification module 126 is to identify, based on a classification of the categorized words 112, the identified rhythm structure 120, and the extracted feature 124, the language 128 of the speech 106. In this regard, the rhythm, features, explicit words, etc., may be classified based on a dictionary (or database) of frequently used words which has been created a-prior for different languages.

[0042] At block 206, the classification may be performed based, for example, on Bayes classifier (looking at posterior probabilities, and creating discrete probability distribution functions based on the words being identified and their frequencies).

[0043] Alternatively or additionally, the classification at block 206 may be performed based, for example, on an artificial neural network (ANN) structure. The ANN may employ, for example, one neural network per language or one neural network with multiple output neurons, with one neuron per language. The ANN may be based on an appropriate selection of feature vectors for frequent words corresponding to a specified language.

[0044] Alternatively or additionally, the classification at block 206 may be performed based, for example, on an unsupervised approach such as fuzzy c-means or k-means clustering based on feature vectors being clustered according to the language class.

[0045] Alternatively or additionally, the classification at block 206 may be performed based, for example, on a Hidden Markov Model.

[0046] Alternatively or additionally, the classification at block 206 may be performed based, for example, on a deep-learning neural network employing 2-d data or 3-d data. An example of 2-d data may include time and amplitude of the speech words. An example of 3-d data may include a spectrogram or red, green, blue (RGB) mapped spectrograms.

[0047] For the classification at block 206, there may be one model for performing classification and identification, or a parallel structure employing multiple-models for classification.

[0048] The model(s) with respect to block 206 may be trained to account for speech rate variations as well.

[0049] Figure 3 illustrates an example of language identification using speech and camera input to illustrate operation of the apparatus 100.

[0050] Referring to Figure 3, due to the presence of a camera (e.g., see camera input at 300), after ambient light compensation at block 302, a-priori landmark features from block 304 of the user 108 may be applied to pre-trained models at block 306. The pre-trained models may use anthropometric data based on racial characteristics. Alternatively or additionally, the image of the user 108 may be applied to a pre-trained deep-learning model to classify the user 108 based on ethnicity. In this regard, referring to block 308, the language identification module 126 is to combine the results of the visual demographic classification with speech identification to increase classification rates in the presence of certain confounds. For example, the confounds may include residual ambient noise corrupting the

speech signal, acoustical environment degrading speech quality, infrequently used words in the speech capture, rhythm mis-estimation, etc.).

[0051] For block 306, an example of a machine learning model may include a deep neural network (DNN) such as a long short-term memory (LSTM), or related model. The machine learning model may be trained to discriminate languages, and include the application of pre word or post word segmentation. Facial feature tracking may also be used as an additional modality to reinforce phoneme matching.

[0052] For the camera-based approach of Figure 3, this approach may employ weighting factors to account for the results of blocks 206 and 306 (e.g., see block 308). For example, assuming that the results of the classification at block 206 include a 55% probability, and results of the classification at block 306 include an 80% probability that the user 108 is of a particular ethnicity, then the confidence of the overall result of the classifier may be increased based on the probability determined by the results of block 306. For example, weights may be assigned to the classification results as follows:

$$(W1 \times \text{Classifier for speech} + W2 \times \text{Classifier for image}) / (W1 + W2) \quad \text{Equation (1)}$$

The weights (W) in this regard may be assigned over a period of time or a specified duration of words (e.g., 10 words).

[0053] Figure 4 illustrates an example of feature based training and transfer of a neural network model to illustrate operation of the apparatus 100.

[0054] Referring to Figure 4, according to an example, multiple speech signals may be emulated as being common words. For example, an English word “family” (spoken by a female), and a Russian word “sem’ya” (equivalent to family) spoken by a male as well as a female may be used. In this example, the feature being used, among the various features described herein, may include the fundamental frequency f_0 (Hz).

[0055] At block 400 of Figure 4, for a speech signal, training of a machine

learning model may be performed at 402, and authentication based on a speech signal at block 404 may be performed at 406. The intermediate blocks between block 400 to block 402 may include Hamming window and frame-hop, feature extraction, and feature normalization over the duration of speech. Similarly, the intermediate blocks between block 404 to block 406 may include Hamming window and frame-hop, feature extraction, and feature normalization over the duration of speech.

[0056] Figure 6 illustrates English emulated "common" spoken word (family) by a female, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the apparatus 100. Figure 7 illustrates Russian emulated "common" spoken word (sem'ya for family) by a male, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the apparatus 100. Figure 8 illustrates Russian emulated "common" spoken word (sem'ya for family) by a female, where the top subplot is the spoken word and the bottom subplot is the fundamental frequency f_0 tracked over the word for a frame size of 512 samples, to illustrate operation of the apparatus 100.

[0057] Referring to Figures 6-8, the fundamental frequency f_0 may be tracked over the duration of the word (the spoken word shown at the top subplot, whereas f_0 shown in the bottom subplot).

[0058] Figure 9 illustrates processed female English tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 6, to illustrate operation of the apparatus 100. Figure 10 illustrates processed male Russian tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 7, to illustrate operation of the apparatus 100. Figure 11 illustrates processed female Russian tracked feature f_0 to remove redundancies (e.g., zeros in the f_0 value), corresponding to Figure 8, to illustrate operation of the apparatus 100.

[0059] Referring to Figures 9-11, the similarities in the female and male Russian spoken f_0 with respect to the envelope may be observed, whereas clear differences in the envelope of the English spoken word f_0 may be observed.

[0060] The envelope may then be used to train a machine learning model as shown at 402. Specifically, a neural network model, as shown in Figure 5, employing a feed-forward artificial neural network (ANN) with N_h hidden layers (e.g., $i = \{1, \dots, N_h\}$ and m_i neurons for hidden layer N_i) may be trained on such feature vectors or their transforms (including envelopes) to generalize and classify according to the languages, by exploiting the similarities of the commonly spoken word in a given language (between male and females), and dissimilarities due to accent and differences due to the commonly used word translations. In this example, the training exemplars may include the fundamental frequency f_0 over the duration of the spoken word (e.g., the boundaries of the word may be determined when the threshold of the amplitude of the signal falls below a certain value factoring any residual/ambient noise) for the various languages.

[0061] For each language, an output neuron may be trained to identify the most common word in that language with a categorical label for that language, or a class value of 1. The neurons associated with the remaining languages in the ANN model may be trained to report an output of -1. The ANN may be trained with rate changes of the commonly spoken word (for both male and female in a given language) using, for example, a vocoder. The ANN model may also be trained with various accents.

[0062] The training algorithm may be gradient descent based, Jacobian, or Hessian matrix based. For example, the Levenberg-Marquardt approach may be used to train the model in a backpropagation mode, where the weight vectors of the neural network may be adapted to minimize the mean-square output classification error, via an approximation to the Hessian matrix of errors (e.g., second-order derivatives), according to:

$$\underline{w}_{k+1} = \underline{w}_k - (\underline{J}^T \underline{J} + \mu \underline{I})^{-1} \underline{J}^T \underline{e} \quad \text{Equation (2)}$$

For Equation (2), $\underline{J}$ may represent the Jacobian matrix that contains first derivatives of the network errors with respect to the weights and biases, and $\underline{e}$ may represent a vector of network errors. The Jacobian matrix may be determined through a backpropagation technique, as opposed to determining the Hessian matrix.

[0063] The size of the input layer, for example, for the ANN may be determined by the product of the number of features tracked and the duration of the feature (or the duration of the envelope of the feature being tracked).

[0064] An example of a machine learning model may include a DNN such as a LSTM, or related model. The machine learning model may be trained to discriminate languages, and include the application of pre word or post word segmentation. Facial feature tracking may also be used as an additional modality to reinforce phoneme matching.

[0065] Figures 12-14 respectively illustrate an example block diagram 1200, an example flowchart of a method 1300, and a further example block diagram 1400 for language identification with speech and visual anthropometric features. The block diagram 1200, the method 1300, and the block diagram 1400 may be implemented on the apparatus 100 described above with reference to Figure 1 by way of example and not limitation. The block diagram 1200, the method 1300, and the block diagram 1400 may be practiced in other apparatus. In addition to showing the block diagram 1200, Figure 12 shows hardware of the apparatus 100 that may execute the instructions of the block diagram 1200. The hardware may include a processor 1202, and a memory 1204 (i.e., a non-transitory computer readable medium) storing machine readable instructions that when executed by the processor cause the processor to perform the instructions of the block diagram 1200. The memory 1204 may represent a non-transitory computer readable medium. Figure 13 may represent a method for language identification with speech and visual anthropometric features, and the steps of the method. Figure 14

may represent a non-transitory computer readable medium 1402 having stored thereon machine readable instructions to provide language identification with speech and visual anthropometric features. The machine readable instructions, when executed, cause a processor 1404 to perform the instructions of the block diagram 1400 also shown in Figure 14.

[0066] The processor 1202 of Figure 12 and/or the processor 1404 of Figure 14 may include a single or multiple processors or other hardware processing circuit, to execute the methods, functions and other processes described herein. These methods, functions and other processes may be embodied as machine readable instructions stored on a computer readable medium, which may be non-transitory (e.g., the non-transitory computer readable medium 1402 of Figure 14), such as hardware storage devices (e.g., RAM (random access memory), ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM), hard drives, and flash memory). The memory 1204 may include a RAM, where the machine readable instructions and data for a processor may reside during runtime.

[0067] Referring to Figures 1-12, and particularly to the block diagram 1200 shown in Figure 12, the memory 1204 may include instructions 1206 to analyze an input signal 104 to identify a speech 106 of a user 108.

[0068] The processor 1202 may fetch, decode, and execute the instructions 1208 to determine, based on word boundary detection, a plurality of words 112 included in the speech 106.

[0069] The processor 1202 may fetch, decode, and execute the instructions 1210 to categorize the words 112, for example, into a phoneme, a vowel, and/or a consonant, to generate categorized words 116.

[0070] The processor 1202 may fetch, decode, and execute the instructions 1212 to identify a rhythm structure 120 of the speech 106.

[0071] The processor 1202 may fetch, decode, and execute the instructions

1214 to extract a feature 124 of the speech 106.

[0072] The processor 1202 may fetch, decode, and execute the instructions 1216 to identify, based on a classification of the categorized words 112, the identified rhythm structure 120, and the extracted feature 124, a language 128 of the speech 106.

[0073] The processor 1202 may fetch, decode, and execute the instructions 1218 to control processing associated with the speech 106 based on the identified language 128.

[0074] Referring to Figures 1-11 and 13, and particularly Figure 13, for the method 1300, at block 1302, the method may include analyzing an input signal 104 to identify a speech 106 of a user 108.

[0075] At block 1304 the method may include analyzing, by a trained recurrent neural network 134, the speech 106 of the user 108.

[0076] At block 1306 the method may include extracting, based on the analysis of the speech 106 by the recurrent neural network 134, features of the speech 106.

[0077] At block 1308 the method may include identifying, based on a classification of the features by the recurrent neural network 134, a language 128 of the speech 106.

[0078] At block 1310 the method may include controlling processing associated with the speech 106 based on the identified language 128.

[0079] Referring to Figures 1-11 and 14, and particularly Figure 14, for the block diagram 1400, the non-transitory computer readable medium 1402 may include instructions 1406 to analyze an input signal 104 to identify a speech 106 of a user 108.

[0080] The processor 1404 may fetch, decode, and execute the instructions 1408 to determine, based on word boundary detection, a plurality of words 112 included in the speech 106.

[0081] The processor 1404 may fetch, decode, and execute the instructions 1410 to extract a feature 124 of the speech 106 by analyzing a fundamental frequency of a word of the plurality of words 112 by a trained learning model.

[0082] The processor 1404 may fetch, decode, and execute the instructions 1412 to identify, based on a classification of the extracted feature 124, a language 128 of the speech 106.

[0083] The processor 1404 may fetch, decode, and execute the instructions 1414 to control processing associated with the speech 106 based on the identified language 128.

[0084] What has been described and illustrated herein is an example along with some of its variations. The terms, descriptions and figures used herein are set forth by way of illustration only and are not meant as limitations. Many variations are possible within the spirit and scope of the subject matter, which is intended to be defined by the following claims -- and their equivalents -- in which all terms are meant in their broadest reasonable sense unless otherwise indicated.

What is claimed is:

1. An apparatus comprising:

a processor; and

a non-transitory computer readable medium storing machine readable instructions that when executed by the processor cause the processor to:

analyze an input signal to identify a speech of a user;

determine, based on word boundary detection, a plurality of words included in the speech;

categorize the words into at least one of a phoneme, a vowel, and a consonant;

identify a rhythm structure of the speech;

extract a feature of the speech;

identify, based on a classification of the categorized words, the identified rhythm structure, and the extracted feature, a language of the speech; and

control processing associated with the speech based on the identified language.

2. The apparatus according to claim 1, the instructions are further to cause the processor to:

analyze the input signal to identify the speech of the user by performing, based on at least one of a zero-crossing rate and a periodicity analysis, a voice activity detection.

3. The apparatus according to claim 1, the instructions are further to cause the

processor to:

identify the rhythm structure of the speech by
determining a vowel duration of a vowel included in the words, and
analyzing, based on the vowel duration, a variability in a length of the
vowel included in the words.

4. The apparatus according to claim 1, the instructions are further to cause the processor to:

control processing associated with the speech based on the identified language by at least one of

implementing, for the identified language, a speech to text translation,
and

translating the identified language to another language specific to
another user.

5. The apparatus according to claim 1, the instructions are further to cause the processor to:

analyze landmark features of the user to determine a racial characteristic of the user; and

combine results including the language identification and the determined racial characteristic to increase a confidence associated with the identification of the language of the speech.

6. The apparatus according to claim 1, the instructions are further to cause the processor to:

classify, based on an analysis of an image of the user by a pre-trained learning model, the user according to an ethnicity; and

combine results including the language identification and the classified ethnicity to increase a confidence associated with the identification of the language of the speech.

7. The apparatus according to claim 1, the instructions are further to cause the processor to:

extract the feature of the speech by analyzing a fundamental frequency of a word of the plurality of words by a trained learning model.

8. A method comprising:

analyzing, by a processor, an input signal to identify a speech of a user;

analyzing, by a trained recurrent neural network, the speech of the user;

extracting, based on the analysis of the speech by the recurrent neural network, features of the speech;

identifying, based on a classification of the features by the recurrent neural network, a language of the speech; and

controlling processing associated with the speech based on the identified language.

9. The method according to claim 8, wherein analyzing the input signal to identify the speech of the user further comprises:

performing, based on at least one of a zero-crossing rate and a periodicity analysis, a voice activity detection.

10. The method according to claim 8, wherein controlling processing associated with the speech based on the identified language further comprises at least one of:

implementing, for the identified language, a speech to text translation; and
translating the identified language to another language specific to another user.

11. A non-transitory computer readable medium having stored thereon machine readable instructions, the machine readable instructions, when executed, cause a processor to:

analyze an input signal to identify a speech of a user;
determine, based on word boundary detection, a plurality of words included in the speech;
extract a feature of the speech by analyzing a fundamental frequency of a word of the plurality of words by a trained learning model;
identify, based on a classification of the extracted feature, a language of the speech; and
control processing associated with the speech based on the identified language.

12. The non-transitory computer readable medium according to claim 11, wherein the instructions are further to cause the processor to:

categorize the words into at least one of a phoneme, a vowel, and a consonant;
identify a rhythm structure of the speech; and
identify, based on a classification of the categorized words, the identified rhythm structure, and the extracted feature, the language of the speech.

13. The non-transitory computer readable medium according to claim 11, wherein the instructions are further to cause the processor to:

control processing associated with the speech based on the identified language by at least one of

implementing, for the identified language, a speech to text translation, and

translating the identified language to another language specific to another user.

14. The non-transitory computer readable medium according to claim 11, wherein the instructions are further to cause the processor to:

analyze landmark features of the user to determine a racial characteristic of the user; and

combine results including the language identification and the determined racial characteristic to increase a confidence associated with the identification of the language of the speech.

15. The non-transitory computer readable medium according to claim 11, wherein the instructions are further to cause the processor to:

classify, based on an analysis of an image of the user by a pre-trained learning model, the user according to an ethnicity; and

combine results including the language identification and the classified ethnicity to increase a confidence associated with the identification of the language of the speech.

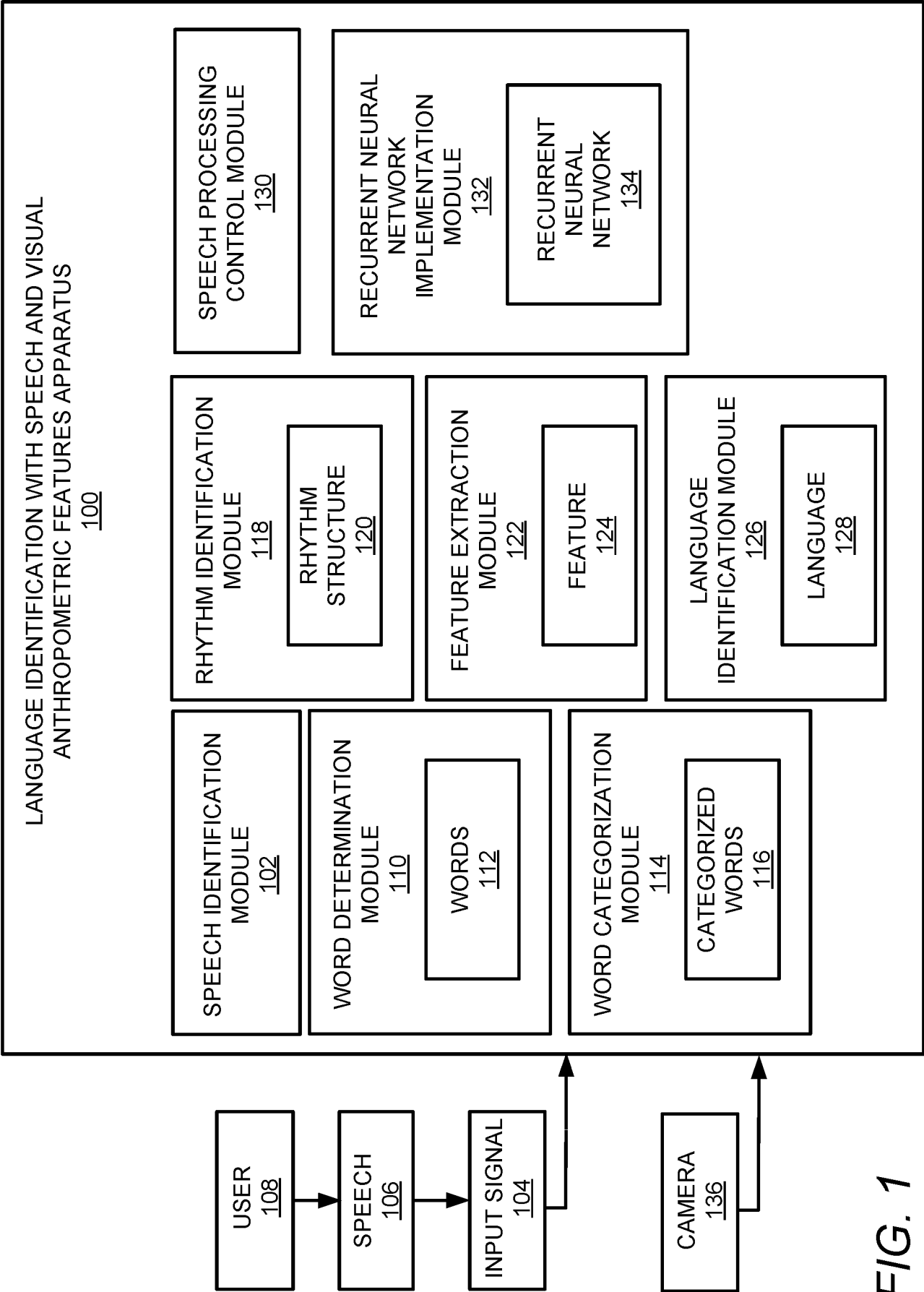


FIG. 1

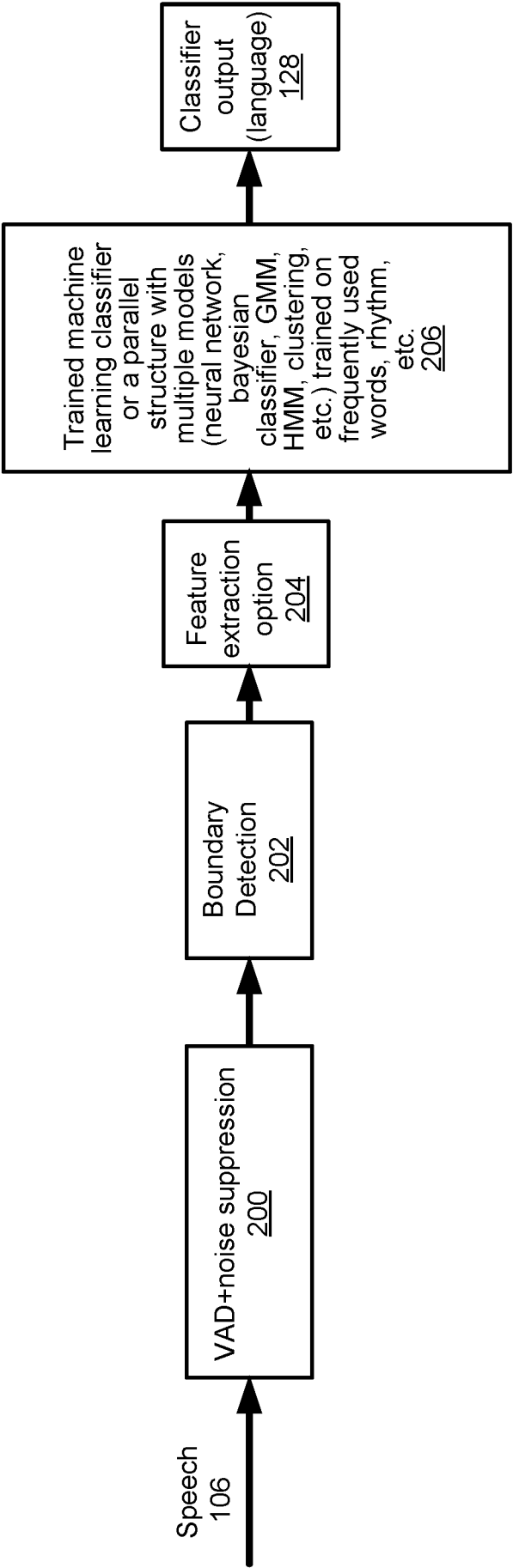


FIG. 2

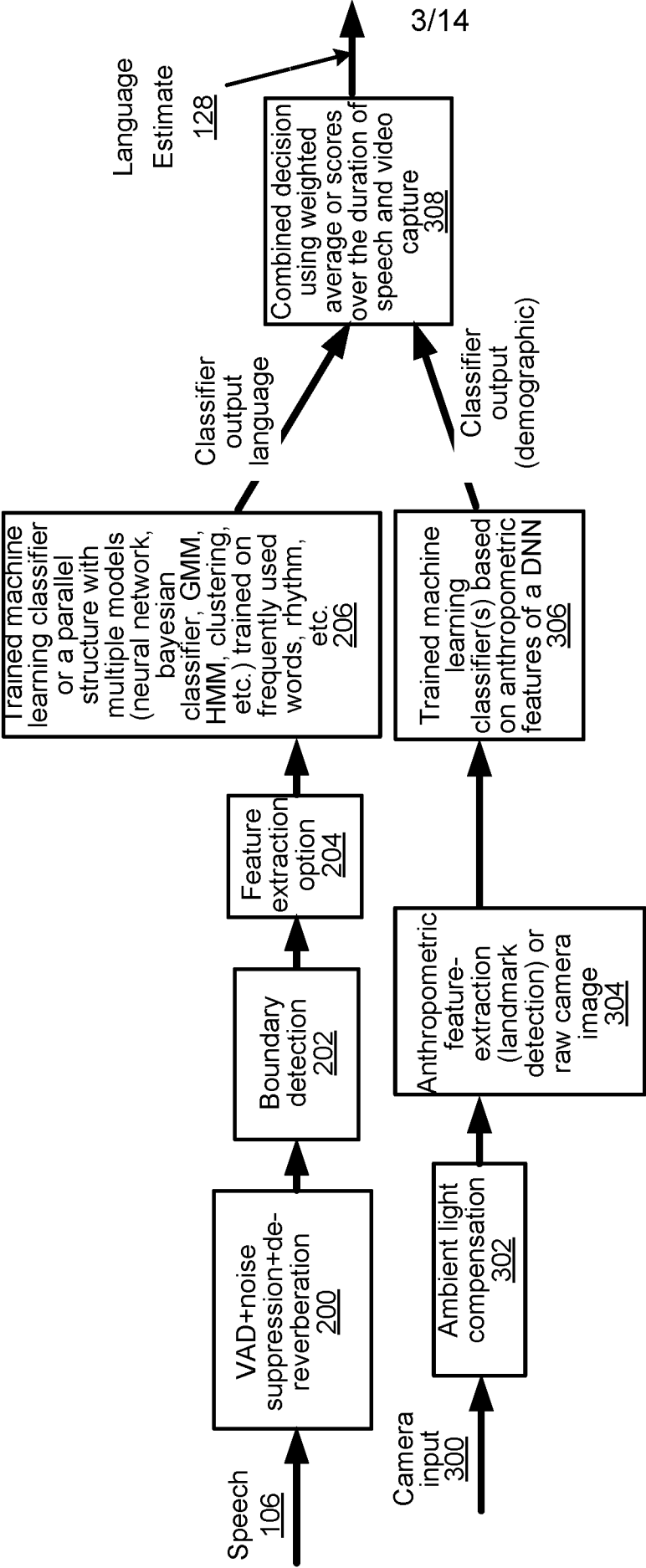


FIG. 3

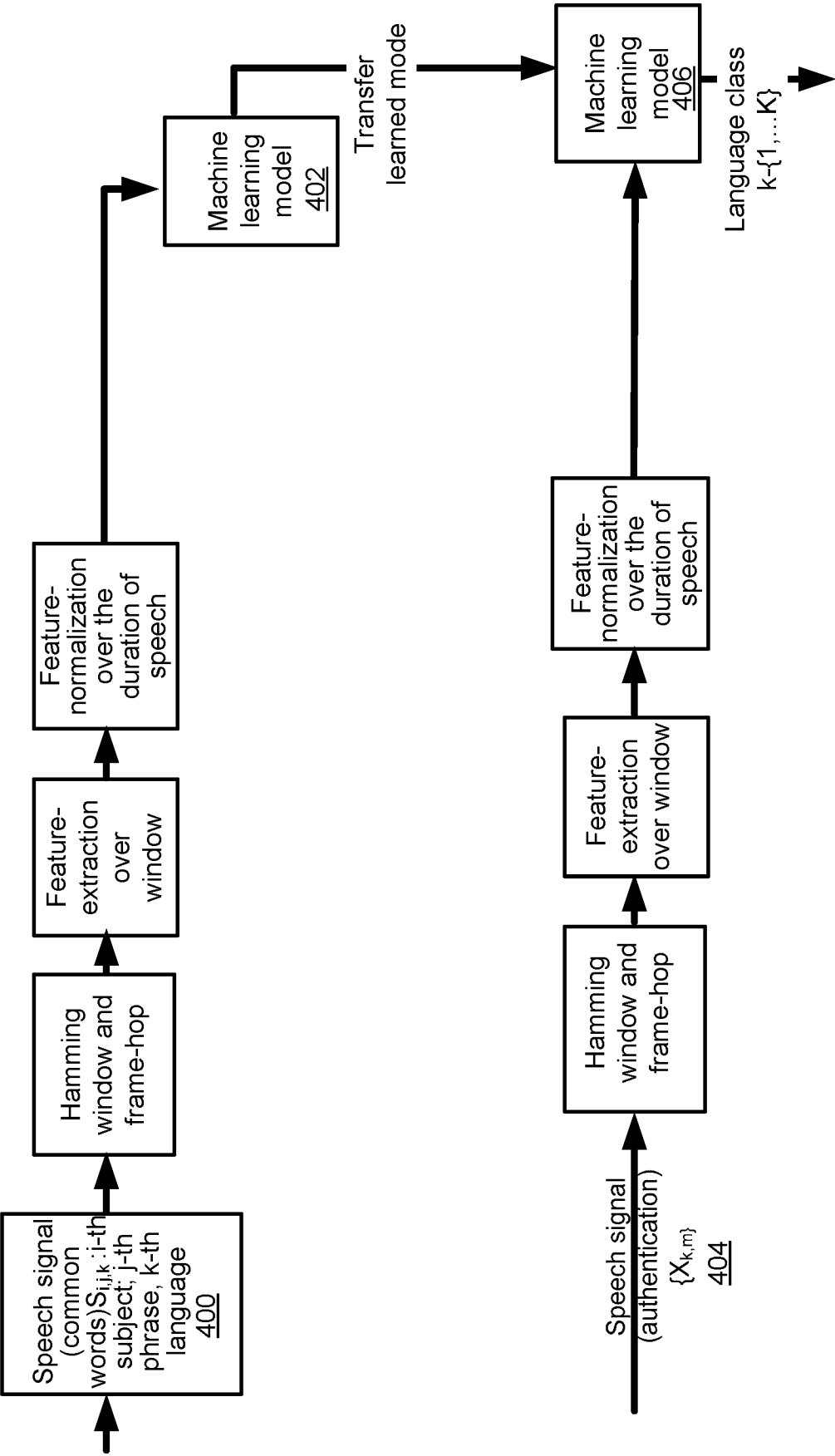


FIG. 4

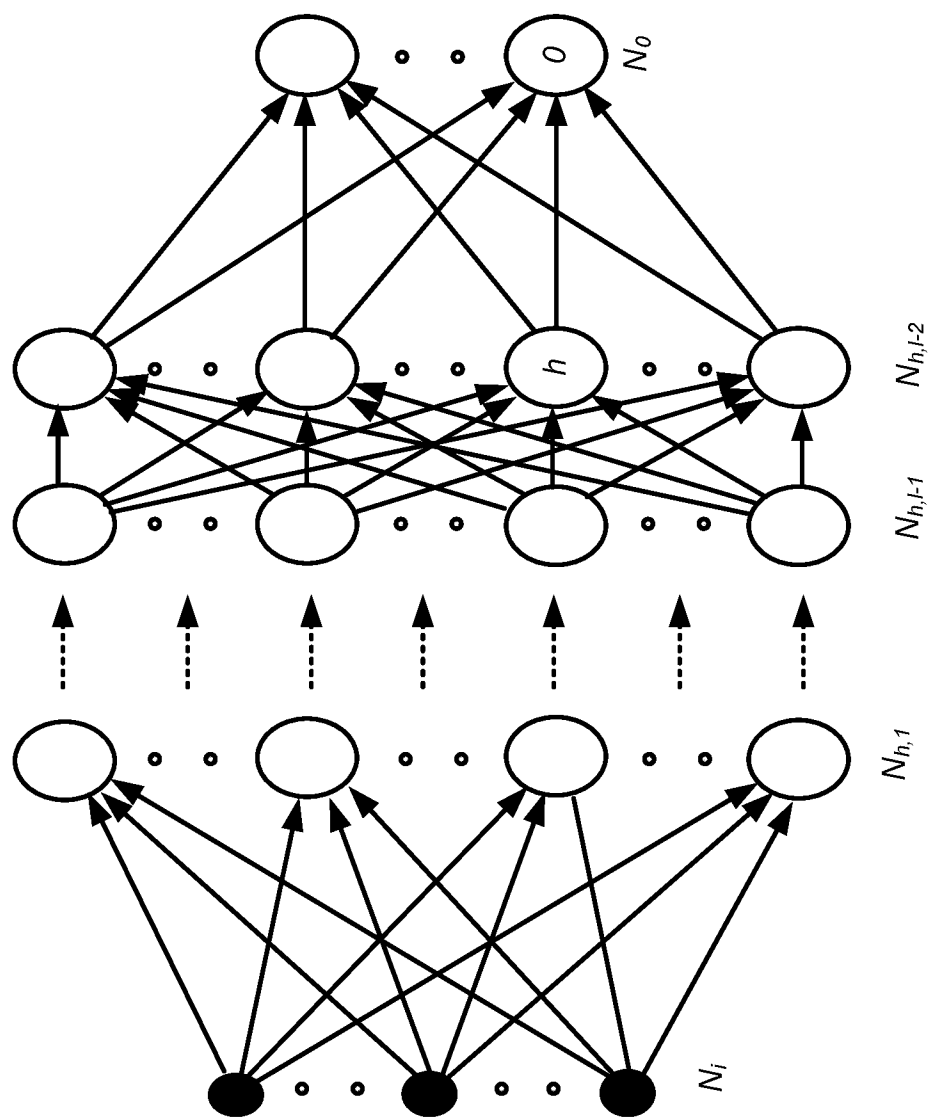


FIG. 5

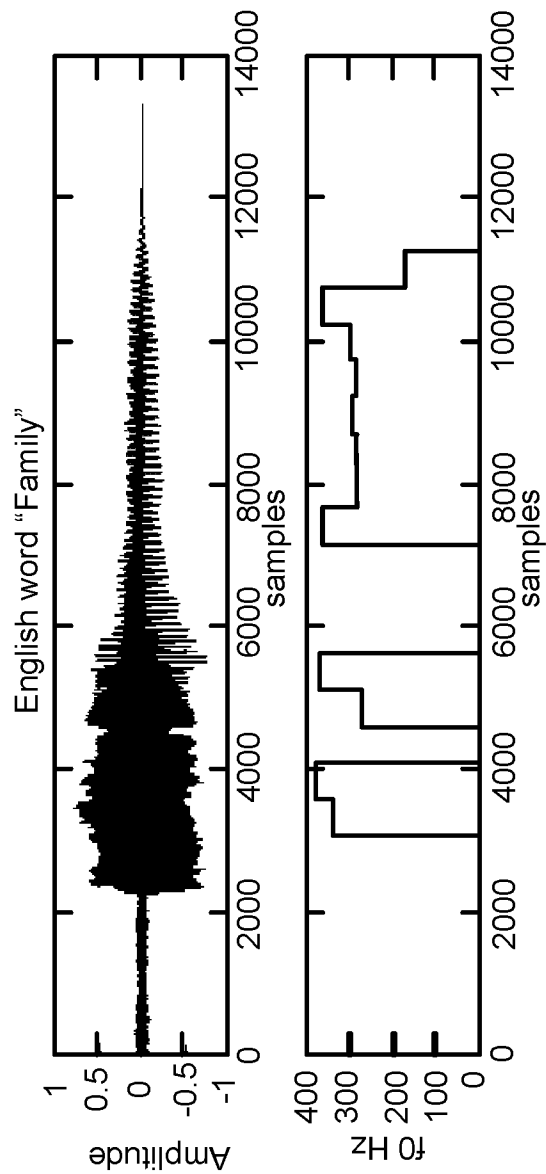


FIG. 6

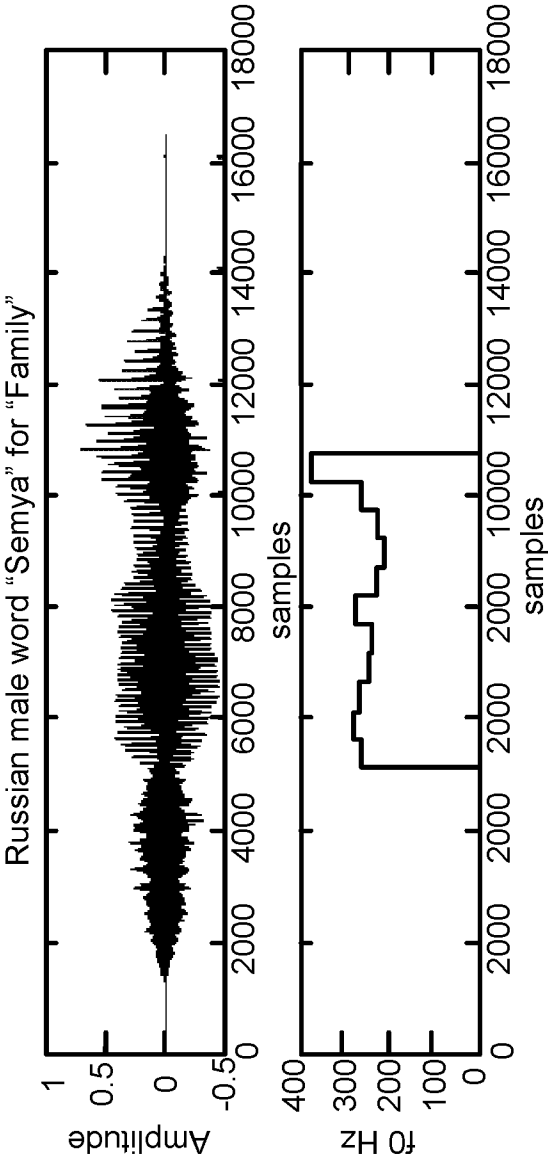


FIG. 7

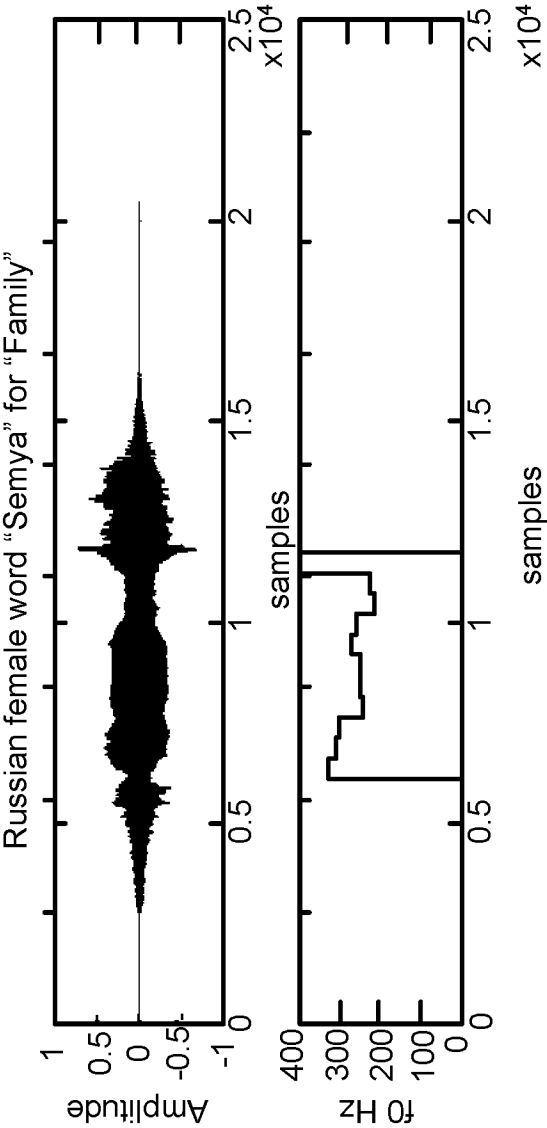


FIG. 8

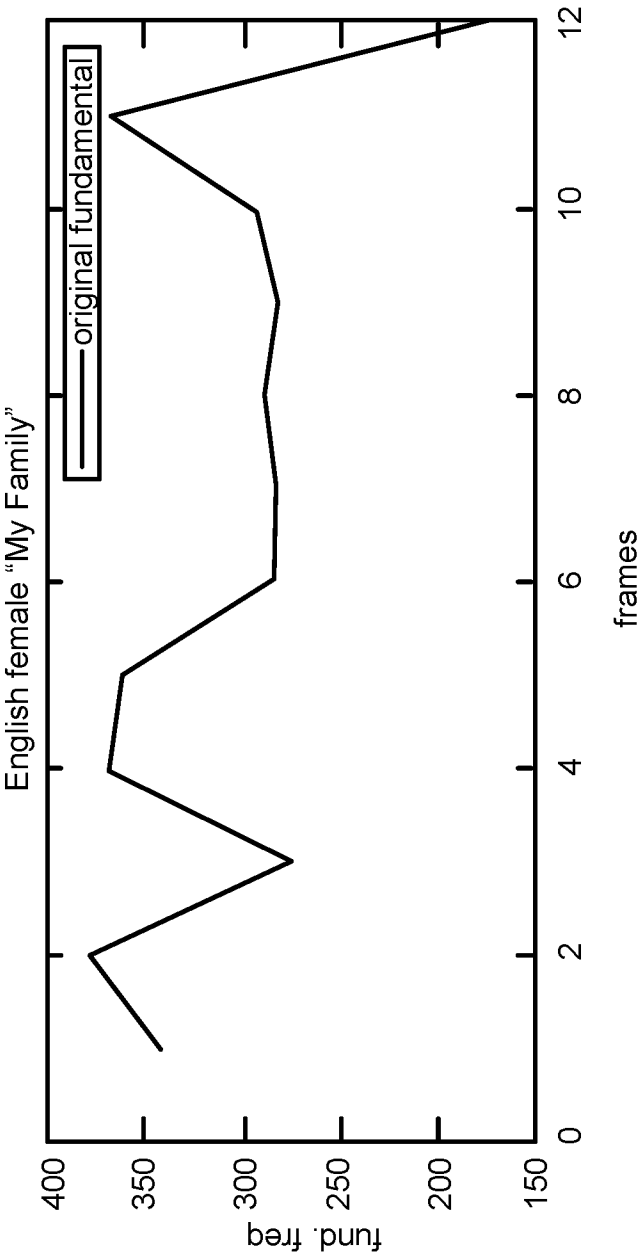


FIG. 9

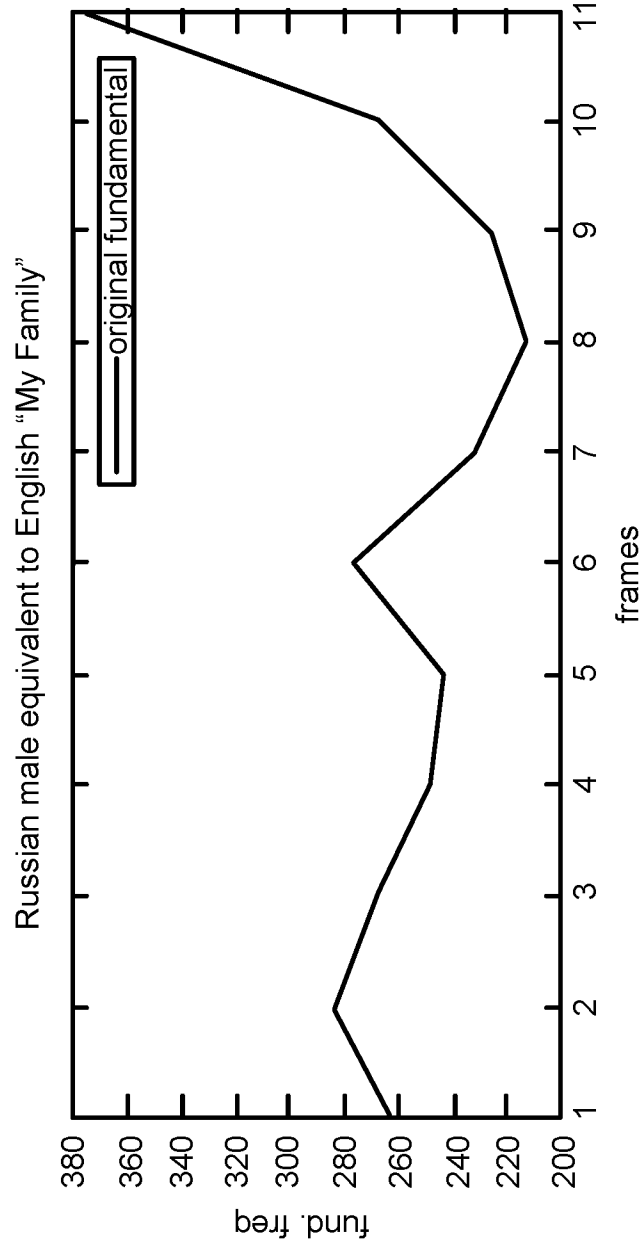


FIG. 10

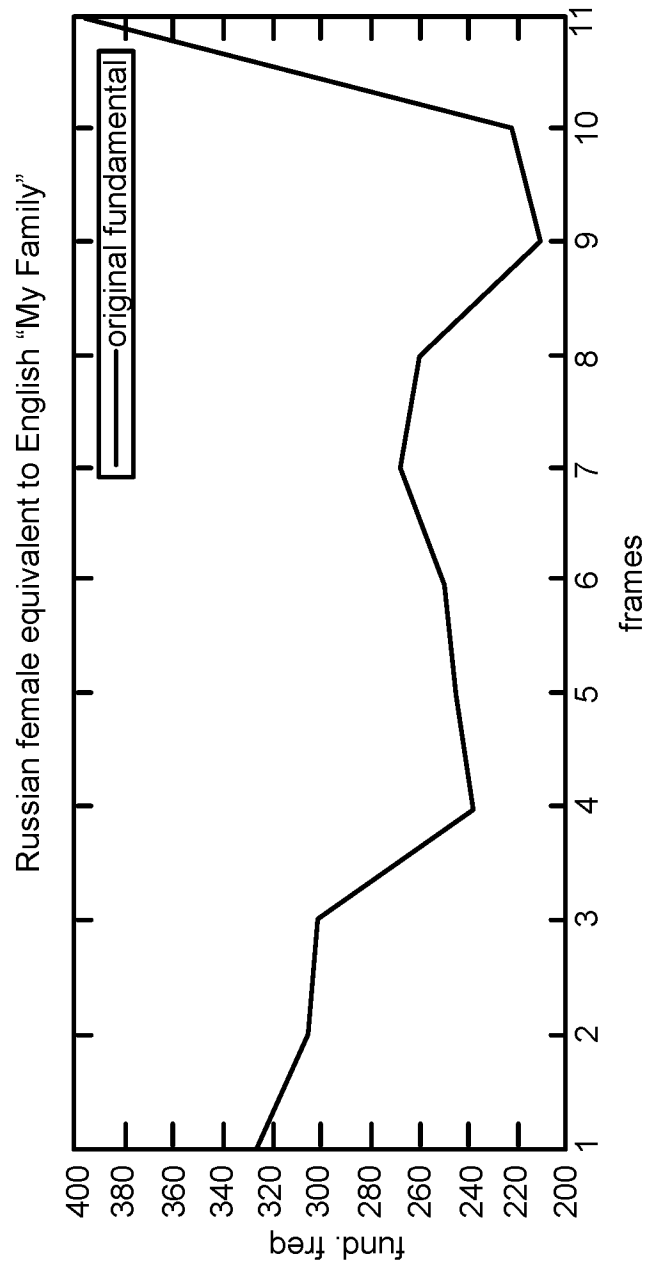


FIG. 11

12/14

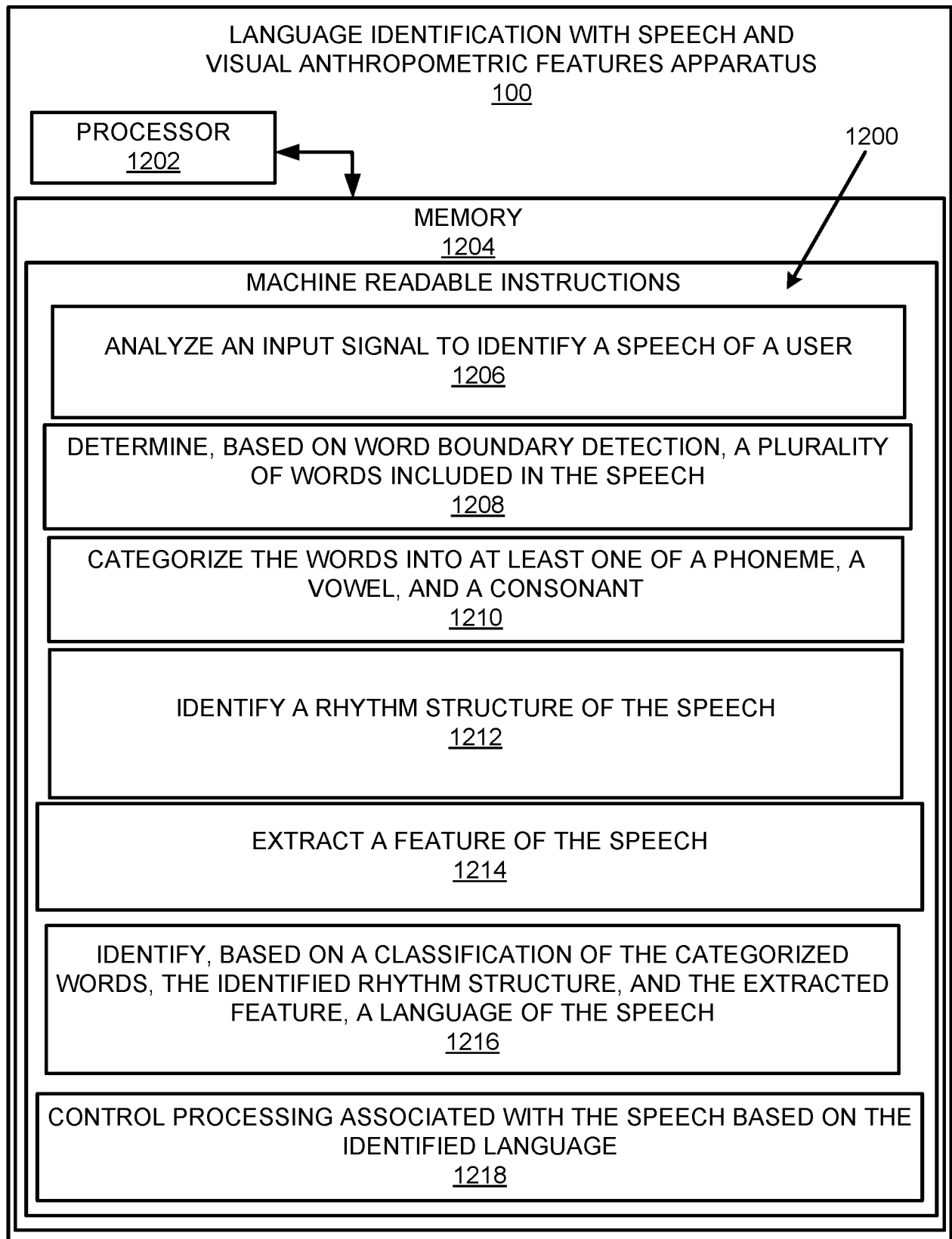
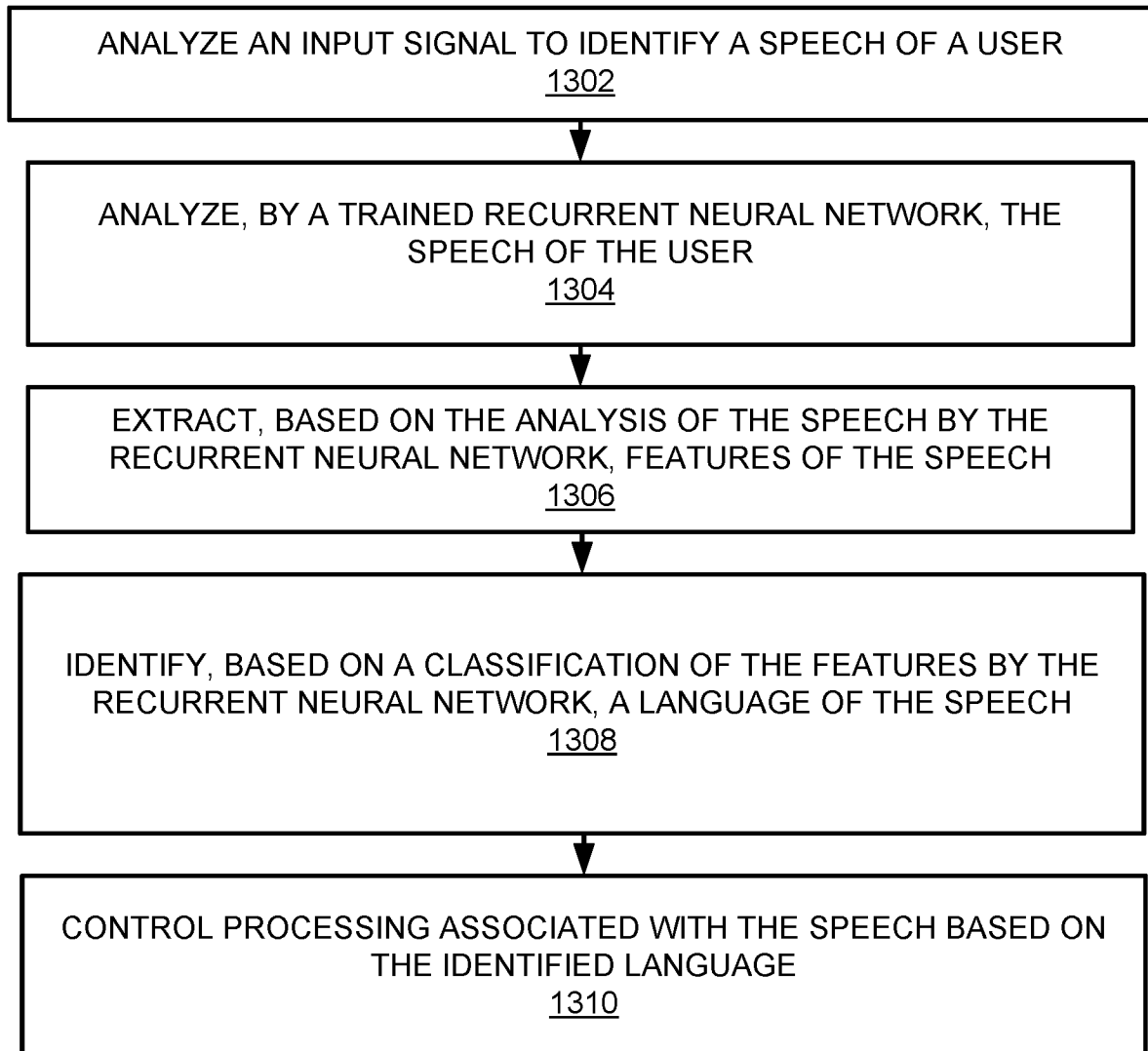
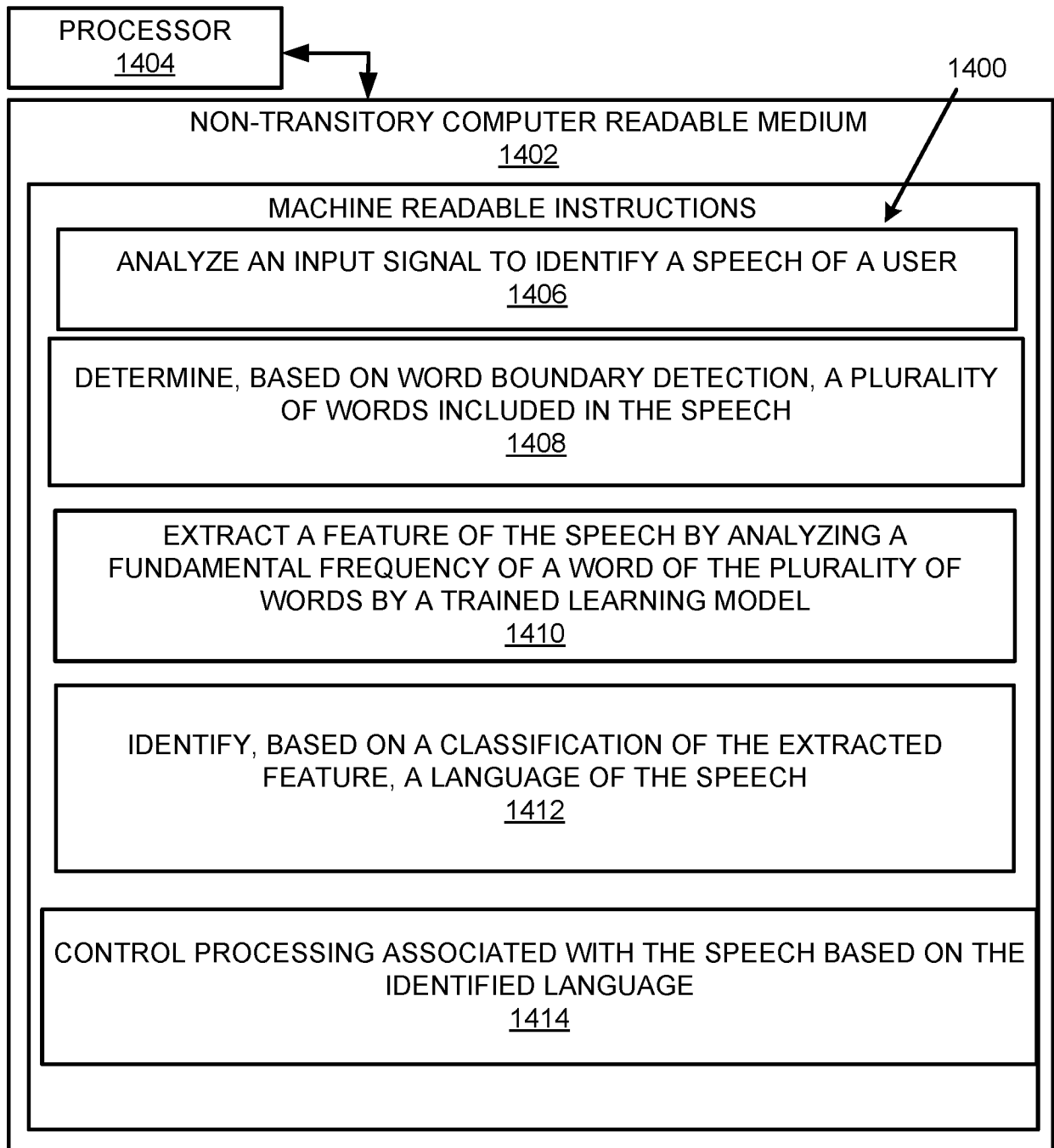


FIG. 12

13/13

1300*FIG. 13*

14/14

*FIG. 14*

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 2017/043765

A. CLASSIFICATION OF SUBJECT MATTER				
<i>G10L 15/16 (2006.01)</i>				
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 15/02, 15/04, 15/06, 15/16, 3/0489, G06F 15/00, 15/16				
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)				
PatSearch (RUPTO internal), Esp@cenet, PAJ, USPTO, Information Retrieval System of FIPS				
C. DOCUMENTS CONSIDERED TO BE RELEVANT				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
X	US 2013/0226583 A1 (AUTONOMY CORPORATION LIMITED) 29.08.2013, paragraphs [0014], [0019], [0023], [0025], [0027], [0029], [0042], [0047], [0054], [0076], [0077], claim 17	1-15		
A	US 2002/0086269 A1 (ZEEV SHPIRO) 04.07.2002	1-15		
A	US 2003/0229497 A1 (LESSAC TECHNOLOGY INC.) 11.12.2003	1-15		
A	US 2007/0260461 A1 (LESSAC TECHNOLOGIES INC.) 08.11.2007	1-15		
☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.				
* Special categories of cited documents: <table border="0" style="width: 100%;"> <tr> <td style="vertical-align: top; width: 50%;"> "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document 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 </td> <td style="vertical-align: top; width: 50%;"> "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 </td> </tr> </table>			"A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document 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
"A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier document 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		
24 January 2018 (24.01.2018)		29 March 2018 (29.03.2018)		
Name and mailing address of the ISA/RU: Federal Institute of Industrial Property, Berezhkovskaya nab., 30-1, Moscow, G-59, GSP-3, Russia, 125993 Facsimile No: (8-495) 531-63-18, (8-499) 243-33-37		Authorized officer V. Volodin Telephone No. (495)531-64-81		